ED 030 303

EM 007 200

By-Ash, Philip
The Relative Effectiveness of Massed Versus Spaced Film Presentation. Rapid Mass Learning. Technical Report.
Pennsylvania State Univ., University Park. Coll. of Education.
Spons Agency-Office of Naval Research, Port Washington, N.Y. Special Devices Center.
Report No-SDC-269-7-3
Pub Date 30 Jun 49
Note-85p.
EDRS Price MF-$0.50 HC-$4.35
Descriptors-Audiovisual Aids, Audiovisual Communication, *Audiovisual Instruction, Film Production, Instructional Aids, *Instructional Films, Instructional Improvement, Instructional Technology, *Mass Instruction, *Military Training, Teaching Techniques, Training Techniques

In presenting material to be learned in a film, is a single, long session, dealing with the subject in depth, as effective as the same content divided into several short sessions? In other words, is a long presentation more tiring than a short one? Groups of psychology students and Navy recruits were given equivalent amounts of instruction time, but according to different protocols--either massed presentation or spaced presentation. For each of the four film series used, the learning was very significant, but the difference between the massed and spaced presentations, as measured by total scores on the film tests, were no greater than could be accounted for by chance alone. Furthermore, the experimental subjects stated that one mode of presentation was not more effective in maintaining interest than the other. The conclusion drawn is that military training films, presently constituting a twenty-minute aid to lecturers, may be lengthened to an hour and become a more central form of instruction. (BB)

TECHNICAL REPORT - SDC 269-7-3

THE RELATIVE EFFECTIVENESS OF MASSED VERSUS SPACED FILM PRESENTATION

(Rapid Mass Learning)

Pennsylvania State College SDC Human Engineering Project 20-
Instructional Film Research Program Contract N6onr-269, T.O.
30 June 1949 Project Designation NR-781

DEPARTMENT OF THE NAVY OFFICE OF NAVAL RESEARCH

SPECIAL DEVICES

ED030303

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

TECHNICAL REPORT - SDC 269-7-3

THE RELATIVE EFFECTIVENESS OF MASSED VERSUS SPACED FILM PRESENTATION

(Rapid Mass Learning)

Pennsylvania State College
Instructional Film Research Program
30 June 1949

SDC Human Engineering Project 20-E-4
Contract N6onr-269, T.O. VII
Project Designation NR-781-005

Report prepared by:

Philip Ash

FOR THE SPECIAL DEVICES CENTER:

Reviewed for Human Engineering Branch:

J. Gaberman

J. Gaberman, Project Engineer
Code 912

Submitted:

H.C. Knutson

H. C. Knutson
Technical Director

H A Voss

H. A. Voss, Head, Code 912

Approved:

J R Ruhsenberger

J. R. Ruhsenberger, Captain, USN
Commanding Officer and Director

EM007200

FOREWORD

This is a report of the results of an experiment on the problem of the relative effectiveness of massed versus spaced training-film presentations. The research has a direct bearing on the question of how long training sessions of motion picture films should be. The report relates indirectly to the question of optimal length of training films. The results have practical implications for the scheduling and use of instructional and informational films.

The Task Order under which the Instructional Film Research Program has been operating requires that attempts be made to establish the scientific principles which should govern both the production and utilization of films for the purpose of rapid, effective mass training. The research results given in this report by Dr. Philip Ash relate to a few aspects of the problems of effective utilization of motion picture films.

It is clear to those who are familiar with the field that the problems are complex and difficult. Nevertheless, it is believed that Dr. Ash has made a significant contribution. Not only have different classes of populations been tested, but also a variety of films has been used. The results point consistently to the main conclusions.

This final technical report is somewhat condensed from the basic thesis, which was presented during June 1949 in partial fulfillment of the requirements for the Doctor of Philosophy Degree in Psychology. This thesis, with all tables, tests, and schedules, has been microfilmed; copies can be made available to individuals who wish to study the full thesis report.

C. R. CARPENTER, Director
Instructional Film Research Program
The Pennsylvania State College

CONTENTS

Page

SUMMARY. 1

I. INTRODUCTION 4
  Background of the Problem 4
  Statement of the Experimental Problem 7
  Review of Related Research. 8
  General Experimental Design 15

II. THE PSYCHOLOGY CLASSES' EXPERIMENT: DESIGN AND PROCEDURES 16
  Films and Tests 16
  Scheduling. 20
  Procedures Followed 21
  The Experimental Population 24

III. THE PSYCHOLOGY CLASSES' EXPERIMENT: RESULTS . . . 29
  The Ape and Child Series. 30
  The Cat Neurosis Series 36
  Relationship Between Test Performances on the Two Film Series 42

IV. THE NAVY EXPERIMENT: DESIGN AND PROCEDURES. . . . 44
  Films and Tests 45
  Scheduling. 47
  Procedures Followed 49
  The Experimental Population 50

V. THE NAVY EXPERIMENT: RESULTS. 52
  Rules of The Nautical Road Series 53
  Elementary Hydraulics Series. 61

VI. COMBINED RESULTS AND DISCUSSION. 69
  General Results 69
  Specific Results. 71

VII. CONCLUSIONS AND RECOMMENDATIONS. 73
  Conclusions 73
  Applications and Recommendations. 73

ACKNOWLEDGEMENTS 76

REFERENCES 77

TABLES

Page

1. FREQUENCY DISTRIBUTION: RUNNING TIME OF ARMY AND NAVY TRAINING FILMS 6
2. EXPERIMENTAL FILM AND TEST SCHEDULE FOR APE AND CHILD SERIES. 22
3. EXPERIMENTAL FILM AND TEST SCHEDULE FOR CAT NEUROSIS SERIES. 23
4. MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL-COLLEGE GRADE-POINT AVERAGE (GPA) AND FINAL PSYCHOLOGY GRADE (PG), FOR SUBJECTS SEEING THE APE AND CHILD SERIES, GROUPED BY CLASS AND METHOD OF PRESENTATION. 26
5. MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL-COLLEGE GRADE-POINT AVERAGE (GPA) AND FINAL PSYCHOLOGY GRADE (PG), FOR SUBJECTS SEEING THE CAT NEUROSIS SERIES, GROUPED BY CLASS AND METHOD OF PRESENTATION. 27
6. MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR APE AND CHILD SERIES: TOTAL TEST SCORE AND SUBTEST SCORES. 31
7. F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS: APE AND CHILD SERIES TEST. . 32
8. MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: APE AND CHILD SERIES. . 34
9. MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR CAT NEUROSIS SERIES: TOTAL TEST SCORE AND SUBTEST SCORES. 37
10. F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS: CAT NEUROSIS SERIES TEST . 38
11. MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: CAT NEUROSIS SERIES . . 41
12. MEANS FOR, AND CORRELATIONS BETWEEN, APE AND CHILD SERIES AND CAT NEUROSIS SERIES TOTAL TEST SCORES: COMPLETE DATA CASES ONLY. 43
13. FILM PRESENTATION AND TESTING SCHEDULE FOR NAVY REPLICATION 48

TABLES (cont'd)

Page

14. MEANS AND STANDARD DEVIATIONS FOR NAVY GENERAL CLASSIFICATION TEST AND MECHANICAL APTITUDE TEST, AND CORRELATIONS BETWEEN THE TESTS: FOR COMPANIES GROUPED BY FILM SERIES AND METHOD OF PRESENTATION. 51

15. MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR RULES OF THE NAUTICAL ROAD SERIES: TOTAL SCORE AND SUBTEST SCORES 54

16. THE EFFECT OF PREVIOUS KNOWLEDGE ON TEST PERFORMANCE, RULES OF THE NAUTICAL ROAD SERIES: TOTAL TEST SCORE MEANS, STANDARD DEVIATIONS, AND MEAN DIFFERENCES. 55

17. F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS: RULES OF THE NAUTICAL ROAD SERIES TEST. 57

18. MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: RULES OF THE NAUTICAL ROAD SERIES. 59

19. MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR ELEMENTARY HYDRAULICS SERIES: TOTAL TEST SCORE AND SUBTEST SCORES. 62

20. THE EFFECT OF PREVIOUS KNOWLEDGE ON TEST PERFORMANCE, ELEMENTARY HYDRAULICS SERIES: TOTAL TEST SCORE MEANS, STANDARD DEVIATIONS, AND MEAN DIFFERENCES. 63

21. F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS: ELEMENTARY HYDRAULICS SERIES TEST. 65

22. MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: ELEMENTARY HYDRAULICS SERIES 67

SUMMARY

Statement of the Problem

The central experimental problem studied in this report may be stated as follows: Given a body of information that is to be presented by means of films, do people learn more if they are presented with this content in one long film in a single session, or if they are presented with the content broken up into several short units in two or more sessions?

A variety of secondary questions was studied in connection with this central one. Of these, the principal one was: Is there a diminution in interest as film length is increased? If so, what relationship is there between learning and interest measured as a function of film length?

Experimental Procedures

The study involved two independent experiments, one with 11 classes of undergraduate psychology students, the other with 10 companies of Navy recruits.

Psychology classes' experiment. Two film series were used. Each series included four 15-minute, silent, black and white reels. The first series presented a comparative study of maturation and learning in a human infant and a chimpanzee. The second series dealt with the induction of experimental neuroses in cats, and showed methods of curing the neuroses.

Three classes were shown each series in a single hour-long session. Two classes were shown each series in two sessions, two reels per session. The sessions, lasting 30 minutes each, were on alternate days. Two classes were shown each series in four sessions, one reel per session. These sessions, lasting 15 minutes each, were on alternate days.

Immediately at the end of each film session, each group was asked to fill out an interest rating form.

One or two weeks after the film showings for each series, the experimental classes were tested on the film content. At the same time, four classes, which served as control groups, were given the film tests without having seen the films.

Navy recruits' experiment. Two films series were used, each comprising three 15-minute sound reels. One series dealt with "rules of the nautical road." The other concerned principles of elementary hydraulics.

Five companies of recruits were assigned to each series. For each series, two companies saw the three reels in a single 45-minute session. Two companies saw the three reels in three 15-minute sessions, one reel per session. One company took the test without seeing the films. The companies were tested one week after the experimental film showings. Interest ratings were made at the end of each film session.

Combined Results and Discussion

Differences in the presentation methods. For each of the four film series, the differences between the massed and spaced presentations methods, as measured by the total scores on the film tests, were no greater than could be accounted for by chance alone. When the scores for the sub-tests for each reel within each film series were analyzed, in general the same results were found. However, for each series, the difference between the control and experimental groups was large and highly significant.

The interest ratings. In general, variation among the groups during any session was as great as, or greater than, variation between methods. Analysis of the distribution of responses to the individual questions on the interest rating forms failed to show any consistent differences among the presentation methods in student or trainee interest. Finally, the correlations between the interest ratings and the film test scores were about zero.

Conclusions

The principal conclusions are:

1. Training sessions using films may last as long as an hour and still result in significant learning. Long massed film sessions have not been shown to be significantly less effective than short spaced sessions.

2. Within the time limits employed in this study, subjects do not seem to find long film sessions less interesting than short spaced sessions, and the learning accomplished seems to be relatively independent of expressed interest.

Recommendations

From this experiment arise the following practical recommendations:

1. In mass training programs, the scheduling of long film sessions for training purposes should be explored as a means of economizing training time, simplifying scheduling, and utilizing instructors more efficiently.

2. Producers should consider the possible advantages of making single long films, where the material calls for extended treatment, rather than series of short units.

3. Further research is needed to determine what the limits are in lengthening film sessions, what kinds of subject matter can be taught in this concentrated manner, and to what sorts of people.

THE RELATIVE EFFECTIVENESS OF MASSED VERSUS SPACED FILM PRESENTATION

I. INTRODUCTION

Background of the Problem

Current educational practice with respect to the use of the instructional film is based largely upon the premise that the film is an aid to the teacher, rather than an exclusive means of instruction. The school day is divided into periods of 40 or 50 minutes each; if there is to be time for the class preparation, the follow-up discussions, and the activities suggested by the film, an instructional film cannot run more than 15 or 20 minutes (13, p. 17). On the basis of this rationale alone, therefore, there would be scant need to inquire into the possibility of efficient learning from longer instructional film units.

In large mass training programs, however, such as those utilized by war industry training organizations and the Armed Forces, there often arises the question of how long instructional film sessions can last and still be effective (33). In these training programs, lack of instructors and lack of time frequently force an elimination of everything but the core content as embodied in the film. The film may be required to do all of the teaching and the instructor may be replaced by a projectionist. Furthermore it may be that in certain educational situations, particularly at the higher levels, the same use may be made of films.

However, limiting the instructional film to 10 or 15 minutes is not a result of the schedule of the school day alone. Calo (4), McKown (21), Doane (10), Bernard (1), and others have suggested that film sessions must be kept short for a variety of other reasons, such as that in long sessions the learners become sleepy and bored, their attention wanders, or the learners may acquire harmful mental habits and be subjected to "hygienic disadvantages." On the other hand, some have maintained that film sessions may run for several hours before serious adverse effects are noted. The convenience of scheduling long sessions is held to offset any slight disadvantages or losses that might be obtained.

It is interesting to note that although Doane (10) is among those who list as one desirable characteristic of instructional films a limit of one reel in length, he points out that the criticism that instructional films are generally too long is not based on any experimental finding. Furthermore, student evaluation of current educational films presents evidence which is directly contrary to the criticism. During the course of the Motion Picture Project of the American Council on Education, 12,000 student ratings were collected for a sample of 500 films. The most frequently mentioned suggestion which the students made for improving both sound and silent films was that such films be made longer. Twenty-six per cent of the students reported that existing films were too short (13, p. 144).

Current learning theory and experiments in learning fail to provide any definitive answer to the question of how long film sessions may last and still be effective. Little relevant work has been done with highly complex materials in instructional films.

A survey of the practices of instructional film producers, however, suggests that the producers have reached a practical solution satisfactory to themselves. Although there are exceptions, the typical commercially produced film is tailored to fit the standard 400-foot reel which runs for just over 10 minutes.

Furthermore, although the Armed Forces were not bound by educational practice and although they frequently used films with little or no instructor embellishment, the Services produced or had produced films which closely approximated these limits. The writer compiled a distribution by length in running time of 1131 Army films and 882 Navy films. The results are reported in Table 1. For each service the mean running time was between 18 and 19 minutes. For each service 56 per cent of the films produced ran for 18 minutes or less. For 89 per cent of the films the running time was less than 30 minutes.

One important consequence of the emphasis on short films, from the point of view of film making, has been the production of series of films, each film a reel long and each covering part of an instructional unit. This practice has been followed by the Armed Forces, the Office of Education, and various commercial producers. At their option, instructors may therefore present each small segment separately, or show all or several of them at a time.

From a practical point of view, then, the question of the relative effectiveness of "long" versus "short" films is of considerable interest. It has a bearing, in the operating training program, on the economics of scheduling and bringing groups of people together; from the production point of view it has relevance to planning the length of films.

It is suggested, however, that an issue more basic than convenience of scheduling or production is involved, namely: Are motion pictures intrinsically different in

TABLE 1

FREQUENCY DISTRIBUTION:
RUNNING TIME OF ARMY AND NAVY TRAINING FILMS

Running Time	Army Films Number	Army Films Per cent	Navy Films Number	Navy Films Per cent
1 - 6 minutes	41	3.63	48	5.45
7 - 12 minutes	299	26.44	192	21.77
13 - 18 minutes	294	26.00	270	30.61
19 - 24 minutes	224	19.80	200	22.68
25 - 30 minutes	144	12.73	74	8.41
31 - 36 minutes	57	5.03	53	6.01
37 - 42 minutes	34	3.00	23	2.61
43 - 90 minutes	38	3.37	22	2.46
Total	1131	100.00	882	100.00
Mean running time	18.99 minutes		18.17 minutes	
Standard deviation	10.69 minutes		8.94 minutes	

their teaching characteristics from lectures or other instructional methods? The raising of the question of film length points to this issue, for practically no educational theorist has proposed that, lest the capacity of the learners be exceeded, class room lectures be reduced to 10 minutes. Even in the most "non-participating" situation - as, for example, in some of our larger schools and colleges where lectures are delivered over public address systems - no one seems to have contemplated limiting the instructional period to much less than an hour.

Current teaching practice, of course, does not constitute adequate evidence for the solution of the problem of determining the optimum length of training film sessions. It may be that films and teachers have quite different effects on learners; or that teaching practice is itself at fault; or finally, that teaching practice is essentially correct and learners can be "safely" exposed to films for periods as long as lectures.

Statement of the Experimental Problem

The experimental problem posed for investigation in this study may be stated as follows.

When instructional films are employed as an exclusive means of teaching (i.e., without instructors, previous preparation, or follow-up discussion), what is the relative effect on measured learning of presenting a standard one-hour film unit in each of the following ways?

1. In one one-hour period (massed presentation method).

2. In two or more equally spaced periods, each lasting a fraction of the hour and including a suitable sub-unit of the hour teaching unit (spaced presentation method).

Secondary questions, some of which emerged as a result of the conduct of the experiment or were suggested by the data, include:

1. What is the relationship between the amount of learning and the subjects' interest in the film series?

2. What is the effect of the possession of previous knowledge on learning from massed, as opposed to spaced, presentation?

3. To what extent do relative differences, if any, persist or change as the retention period is increased?

It should be pointed out that, although this experiment is limited to the situation in which the film is the sole medium of teaching, there does not seem to be a valid reason for believing that the relative efficiency of the presentation methods would be substantially changed if they were used in conjunction with a more conventional technique of film utilization.

Review of Related Research

Except for a single short report of a small study made during World War II by the Morale Services Division of the Army Service Forces, War Department (33), a search of the literature fails to reveal any investigations directly bearing upon the problem which is the subject of this experiment.

However, three lines of study contingent to the problem of the optimum length of training film sessions may be noted. These include:

1. Research on the effectiveness of films in comparison with other training media.

2. Investigations of part-whole learning and massed-distributed practice.

3. Studies dealing with the length of class periods, particularly in the secondary school and college.

These areas will be reviewed briefly first, and then the Army study will be discussed.

Learning from films. Since in the present experiment motion pictures are used as "total teaching procedures" without the aid of instructors and without prior preparation or subsequent discussion, it is pertinent to inquire whether in fact this is a realistic, if not an entirely usual, procedure. As VanderMeer (34) has pointed out, while it has rarely been proposed that films could carry the whole burden of instruction, "Nevertheless, under conditions which may be specified during national emergencies when rapid mass training is required....it may be necessary to utilize....[films]...as a relatively exclusive means of instruction." The question is therefore relevant not only to the experimental design of this particular investigation, but also to the probability of occurrence of teaching situations for which the findings of this investigation might have some degree of pertinence.

Fortunately, in spite of the often-expressed insistence upon the primary role of the instructional film as an aid to the classroom lecture, the motion picture research literature is replete with instances in which the effectiveness of films is explicitly compared with the effectiveness of, among other methods of instruction, the classroom lecture. Of the six studies reported by Devereux (9, pp. 61-100), one by C. C. Clark, of New York University, covered precisely this point. Three equated groups of college students were taught by a series of sound films a series of silent films and a series of classroom lecture demonstrations, respectively. The general conclusion that seemed warranted by the data was that there was no significant difference in the efficiency of the three methods, as measured by subject matter tests.

In a review of the literature, Hoban (8, pp. 334-361) cited a wide variety of studies in which films were compared with lectures and classroom demonstrations, either incidentally or as the central problem of the study. In none of these studies was there evidence to indicate that films were significantly inferior, and occasionally they were found to be significantly better than lectures or demonstrations.

Among the more recent investigators, Jayne (14) compared the factual learning from lectures of one group of freshman students with the learning from silent films of another group. The subject was general science. Jayne found that although the immediate gains from the lectures were higher than those from the films, these differences became less with the passage of time. Philpott (27) compared the performances of five groups, taught by film only, film plus commentary, slides only, slides plus commentary, and oral lesson, respectively. He found very small differences from the "film only" method. Hall and Cushing (12) employed three methods - lecture, reading assignment, and films - and concluded that the learning effected was a function of the material taught and of the learner. None of the three methods was consistently best or worst. Finally, VanderMeer (34) has reported what appears to be the most extensive investigation on this point. Three hundred ninth grade public school students were taught a semester course in general science by one of three methods: exclusive film instruction, film plus prepared study guides, and "typical instructional methods," respectively. The control, or "typical instructional methods," group was taught by an instructor using text books, demonstrations, lectures, and oral questions and answers. The "film only" group saw the films without discussion, teacher comment, or assigned reading. The "film plus study guide" group saw the films and was given mimeographed study outlines for the films, but was given no other instruction. The study guides were not discussed in class. Analysis of the data suggested that the three methods were about equally effective in teaching the subject matter, as measured by factual learning.

It may be concluded, therefore, that (at least for the imparting of factual knowledge) sole dependence on films as teachers is neither impossible nor unrealistic. The evidence suggests that in mass training programs such a procedure may be both practicable and effective. It would be necessary, of course, to explore its limits in terms of the kinds of subject matter to which it might be applied, and the kinds of learning it might bring about or fail to bring about.

Classical learning experiments. Two concepts current in learning theory seem to be relevant to the question of whether there is any significant difference in the effect on retention of presenting a body of material in films massed as a whole in one session, or distributed in parts over several sessions. These are (1) the concept of *massed versus distributed practice* and (2) the concept of *whole versus part learning*. In point of fact, the present experimental situation is not subsumed readily under either of these concepts, and it is difficult to say which is the more pertinent. This experimental situation may be described as one in which presentation of the whole once in a single (massed) session is compared with presentation of the parts in several (spaced) sessions. However, it seems worthwhile to review at least the major findings and to try to relate them to the present design.

Massed versus distributed practice. The practice required to learn a task may be continuous, without rest intervals, or it may be distributed with rest intervals interpolated at a number of points. The relative advantages of these two procedures have been studied in a wide variety of experiments, which have been thoroughly reviewed (*20*, pp. 119-151; *37*, pp. 211-216). McGeoch (*20*, p. 119) states that "the generalization that some form of positive distribution yields faster learning than does massed practice holds over so wide a range of conditions that it stands as one of our most general conclusions."

The large bulk of the experiments, however, deal with rote-memory tasks calling for the learning of nonsense syllables, codes, word lists, or poetry; or with perceptual-motor tasks such as typing, mirror-drawing, mirror reading, or archery. With respect to *complex meaningful materials* the evidence is not as clear-cut. T. W. Cook (in McGeogh, *20*, p. 126), for example, predicted and found that puzzle solution was favored by massed practice. On the other hand, Austin (in McGeoch, *20*, p. 129) found that, while there were no significant *immediate* differences in retention of prose selections whether they were studied in five single readings spaced at intervals of one or two days, or studied in one session of five readings, when the recall intervals were extended to two weeks and a month the balance of superiority shifted to the distributed readings. Gordon and Clark (in McGeoch, *20*, p. 129) both found the same effect for spaced readings of meaningful material.

While positive distribution or spaced practice has *generally* yielded faster learning, however, the experiments have not been extended to material of the complexity found in motion pictures such as the present experimental film series. Furthermore, and perhaps more important, it is doubtful whether the *procedure* employed in this experiment is sufficiently similar to that employed in investigations of massed versus spaced practice to permit meaningful application of the results.

Whole versus part learning. In the typical part-whole experiment, a comparison is made between learning the material repeated as a whole until some criterion of efficiency is reached, and learning the material divided into two or more parts, each part being repeated until a specified criterion of efficiency is reached. The "whole" and the "parts" have usually been defined on a quantitative basis, e.g., as stanzas of a poem. The relevant literature is reviewed by McGeoch (17, 18, 19, 20) and Woodworth (37, pp. 216-223). Three observations seem pertinent. First, the way in which the concept has been defined and measured has required that the tasks used be simple enough to permit establishing a relatively unequivocal final performance criterion. Therefore, the learning of a complex content, which can usually be measured only by a test that samples items from a wide area, has not been explored. Second, the studies have yielded divergent results, many of which were statistically reliable. Conclusions seem to be limited, therefore, to particular tasks, with learning efficiency measured in particular ways, qualified by the assurance that the subjects were similar, and so forth. Third, in view of the foregoing it seems unlikely that this area of research provides any guide to judging whether there is a significant difference in the effectiveness of a single presentation of a whole film as compared with the presentation of the film in parts spaced over several sessions.

In short, although the two learning concepts referred to have at least a nominal similarity to the experimental design, neither has been used in a setting similar to the present one. Furthermore, it may be the case that where subjects are required to apprehend relationships, to grasp concepts and generalizations, and to be able to recognize rather than recite by rote, the traditional "distribution of practice" and "part-whole" concepts are not applicable.

Classroom practice. It has been suggested that the question of how long films sessions can last and still be effective is akin to the question of how long classes can last and still be effective. At least until a reasonable amount of evidence is available to justify a distinction between learning from films and learning from lectures, it would seem pertinent to examine both the experience and the research bearing on the question of the length of class periods.

Two rather different problems have been investigated in connection with the length of class periods. On the one hand, some workers have studied the relative advantages of different amounts of total teaching time during a semester. In these studies, either the frequency of class periods or the length of class periods was varied. On the other hand, some have investigated the effect of changing the length of the period while total teaching time was held constant.

Those studies in which total time was held constant, while unit time was varied, are relevant to the present inquiry.

Most of the literature on this point is discursive, or descriptive of particular programs, rather than experimental. For example, Clevenger (6) and McMillin (22) advocated that class periods be lengthened (generally from 40 minutes to 60 minutes) because "longer periods save money." Clevenger concluded that no one knew the "best length." Nord (25) compared the eight-period day with the six-period day, and concluded that the two are about equal. Greenley (11) advocated 90-minute periods; he reported that, out of 464 high school students in his school, 38 selected single (45-minute) periods while the remainder selected double (90-minute) periods. Manheimer (23) suggested that high schools reorganize on the basis of longer periods; he pointed out that summer-sessions experience with two-hour periods was eminently satisfactory.

Bruns (3) described the experience of one high school in which a shift was made from a schedule of five one-hour periods per day to a three period day, with two 90-minute periods in the morning and one two-hour period in the afternoon. He reported no experimental findings, but claimed that the students and teachers preferred the longer periods. Kambly (15) compared a one-hour, two semester, course in biology with a two-hour, one-semester course; he found "no differences."

Finally, Stewart (30) reported what appears to be the most extensive investigation in this area. He compared the relative effect of lengthening class periods and increasing total time, and also the effect of lengthening class periods with total time held constant. It is the latter part of the experiment which is pertinent here. One hundred and eighty tenth year high school pupils were divided into two equated groups of 90 each. All pupils studied four subjects during a school semester of 12 weeks. In the "regular" group the students carried the four subjects concurrently, for periods of 40 minutes daily for the 12 weeks. The "concentration group" was divided into subgroups to control subject order effects; each subgroup carried two of the subjects during the first six weeks in class periods of 80 minutes daily, and the other two subjects during the second six weeks in periods also of 80 minutes daily. Thus, while total class time was held constant the "concentration group" was taught each subject in a time span half as long as that required for the "regular group."

At the end of the experimental period, standardized achievement tests in the four subjects were administered. In every subject and on every test the concentration group's performance was better than the regular group's performance, and for eight of the 12 tests the ratio of the difference in means to the probable error of the difference was above 4.0 (30, p. 27).

This experiment, therefore, as well as the more or less adequately documented opinions of educators, suggests that efficient learning may be anticipated, at least in the classroom, from a highly concentrated presentation of the subject in periods lasting substantially longer than one hour. The evidence suggests, in fact, that concentrated attention to a few subjects is better than more dispersed attention to several.

The Army study. The specific question of the optimum length of film sessions has been investigated, as far as is known, in only one previous study. This was an investigation made by the Research Branch, Morale Services Division, Army Services Forces (33).

Following is a description of the experimental design employed and conclusions reached:

> "Two standard training films on First Aid (TF 8-33 and TF 8-150), selected for the experiment because of being approximately equal in length and difficulty, were shown to two groups of men. The first group of 350 men were shown both films consecutively in a session lasting about an hour. (This group will be referred to as the Long Session Group.) A second group of 250 men from the same IRTC (the Short Sessions Group) were carefully matched against the first group with respect to intelligence, education and other relevant factors. These men were shown TF 8-33 at a half-hour session in the morning. In the afternoon they were shown TF 8-150 at another half-hour session.
>
> "No difference was found between the Long Session and Short Sessions group in the average percentage of new material imparted by the film which was shown first. However, a significant difference was found between the groups in the amount of new material learned from the film which was shown last.
>
> "It was found upon further study that almost all of the differences between the two groups were accounted for by the slower learners within the two groups. (For the purpose of this analysis, all the men were divided into rapid learners--AGCT classes I and II--and slow learners--AGCT classes III and IV.)"

It was reported that, for the second film, rapid learners in the long session learned 45 per cent new material. Slow learners in the long session, on the other hand, learned only 27 per cent new material. In the short sessions rapid learners learned 46 per cent new material and slow learners learned 35 per cent.

Although the results of this study seem to suggest that spaced film sessions have the advantage, certain questions of design and data analysis tend to weaken this conclusion.

In the first place, it is at least not clear from the statement of the design that the retention test was fairly spaced from the experimental sessions. Let us suppose that the long session group was shown both films in the morning, the short sessions group was shown one in the morning and one in the afternoon, and both groups were tested the following morning. This arrangement seems not unlikely on the basis of the description available. In this case, the short sessions group would have enjoyed about a four-hour retention advantage for the second film. This is a sixth of the 24 hour span, and could well have a significant effect on performance. For such short intervals the slope of the retention curve tends to be still quite steep (e.g., cf. 37, p. 53).

In the second place, the results are expressed only in terms of per cent indices. Per cent indices are at best uncertain statistics. They do not necessarily reflect the absolute magnitude of the original scores. In general, regardless of the statistical significance of a difference in mean scores, these indices will tend to yield larger percentage differences at one end of the performance scale, and smaller differences at the other end.[1]

In view, therefore, both of the questionable character of the experimental design and the ambiguity of the statistics, it is doubtful that much confidence can be placed in the findings reported.

This survey of literature dealing with problems contingent to the present one suggests that research on these problems has at most provided very general and somewhat ambiguous conclusions. It has been demonstrated that films can be used effectively as total teaching devices. There is a considerable body of opinion and some experimental evidence to indicate that a small number of long class sessions led by instructors are, if anything, better than a large number of short sessions. One small study suggests, without arousing conviction, that long film sessions are not as effective as short spaced sessions for slow learners.

---

[1] See microfilm of the original dissertation: footnote 1, page 21. That footnote embodies a statistical critique of the per cent indices used.

General Experimental Design

The research to be reported consists of two relatively independent experiments. It has not been thought feasible to deal with these experiments simultaneously. Therefore, it may be profitable to review the basic experimental design common to both, and to point out the essential differences between them.

In the first experiment, two series of four 15-minute films were shown, a month apart, to the introductory psychology classes at The Pennsylvania State College. Three methods of presentation were employed: a massed presentation method using a single session lasting one hour and including all four films, a spaced presentation method using two sessions lasting 30 minutes each and including two films per session, and a spaced presentation method using four sessions lasting 15 minutes each and including one film per session.

In the second experiment, two series of three films were shown to Navy apprentice seamen at the Great Lakes Naval Training Center, Great Lakes, Illinois. In this experiment, the two series were shown concurrently, to different groups of men. Two methods of presentation were employed: a massed presentation method using a single session lasting 45 minutes and including all three films, and a spaced presentation method using three 15-minute sessions and including one film per session.

In both experiments the subjects rated the experimental films as to interest value, immediately at the end of each session. All groups were tested one week after the experimental presentations, with the exception of three classes which were shown the second psychology series. These classes were tested approximately two weeks later.

For each film series, an analysis was made of the relative efficiency of the methods as measured by the test scores of the experimental group subjects. However, for each series the test was also administered to an appropriate control group which had not been shown the series. The performance of the control group served as a basis for estimating the "absolute" contribution of the films to the subjects' knowledge, irrespective of the method of presentation. This use of the control group has two important advantages (in factual learning experiments) over the use of pre-tests to determine initial status with respect to the experimental content: it avoids sensitizing the group to "fact-quiz" items, and, as a corollary, it tends to maximize inter-group differences. Particularly with tests of low reliability, the "after-only" procedure tends to be the more sensitive.

II. THE PSYCHOLOGY CLASSES' EXPERIMENT: DESIGN AND PROCEDURES

The first experiment was conducted with 11 classes of introductory psychology students as subjects. Seven classes served as experimental groups. These classes were shown two four-reel series of motion pictures. The remaining four classes were used as control groups. These classes took the tests on the films, but did not see the films. Three presentation methods were employed: a one-part method, in which all four reels were shown during a single class period; a two-part method, in which the first two reels were shown in one period, and the second two in a succeeding class period; and a four-part method, in which one reel was shown in each of four periods.

In this chapter, the films and tests, the scheduling, the specific experimental procedures, and the subjects will be described. In the following chapter the statistical results obtained from this experiment will be presented.

Films and Tests

Film criteria. The following criteria were established for the film material to be used in the study:

1. The films should have general technical adequacy, in terms of photography, coverage of teaching content, clarity of presentation and so forth.

2. The units of any one series should be produced with sufficient standardization of treatment to permit combination and smooth transitions.

3. The films should be appropriate for the experimental populations involved.

4. Duplication of material in the units comprising a series should be held to a minimum.

5. If possible, the units of a series should not have a necessary sequence, to permit the study of order effects.

6. The series should consist of four units of approximately equal length, and the four together should run about an hour.

7. The film material should be new to the learners, if possible.

8. The films should be non-dramatic, factual presentations.

When specific films were examined with these criteria in mind, several points became evident. First, in order that the second criterion, that of homogeneous treatment, be met, it became clear that an already existing series would have to be employed. It was not possible to combine independently produced units into a series, and at the same time to avoid duplication and to realize smooth transitions. However, as a result of this conclusion, it was necessary to drop the fifth criterion, relating to effects of order, since existing series almost invariably had a definite sequence. This limitation was not too serious, because it soon became apparent that, even if order could be varied, the large number of groups necessary to vary order of films within a series would not be available.

When it was indicated that the introductory psychology classes at the College would be made available, the film catalogues, particularly that of the Psychological Cinema Register (26), were searched and a large number of films were screened. Several series of films were considered, only to be rejected because they were judged too difficult, too short, too long, or otherwise inappropriate. The two series finally selected seemed to meet all the criteria to a satisfactory degree. These two series each consisted of silent black and white, 16 millimeter films. When run at normal speed (16 frames per second) they took more time than could be allowed. However, viewing tests revealed that they could be run at sound speed (24 frames per second) without loss of visual quality or too hasty presentation of the explanatory titles. Accordingly, in the experiment the films were shown at sound speed. Thus, each film was shown 50 per cent faster than the producer intended, and as a result it was shown in two-thirds of the time usually required.

The films used. The film series used were (1) Dr. W. N. Kellogg's *The Ape and Child Series*, and (2) Dr. Jules H. Masserman's *The Dynamics of an Experimental Neurosis Series*. A brief outline of the film content, based largely on the *Psychological cinema register catalogue* descriptions, follows:

PCR-80-83: *The Ape and Child Series*.

PCR-80: *Some Behavior Characteristics of a Human and a Chimpanzee Infant in the Same Environment* (running time, 14.8 minutes). The general behavior of a normal human infant between the ages of 10 and 14½ months is compared step by step with analogous behavior of his chimpanzee companion between the ages of 7½ and 12 months. Six phases of behavioral development are illustrated, as well as the early effects of human environment upon the ape and some basic differences between the ape and the child.

PCR-81: Comparative Tests on a Human and a Chimpanzee Infant of Approximately the Same Age (running time, 14.5 minutes). The reactions of the human infant are compared with the responses of the chimpanzee to a series of psychological tests. The tests include: handedness, startle reaction time, delayed reaction, cap-on-head tests, rotation tests, and others.

PCR-82: Experiments Upon a Human and a Chimpanzee Infant After Six Months in the Same Environment (running time, 14.5 minutes). Some of the more complex tests solved by the human infant, age 16 to 19 months, and the chimpanzee, age 13½ to 16 months, are demonstrated. Five tests are presented, involving simple perceptual-motor tasks.

PCR-83: Some General Reactions of a Human and a Chimpanzee Infant After Six Months in the Same Environment (running time, 14.4 minutes). "Incidental" or non-experimental behavior of the human infant and the chimpanzee are compared. Nine types of comparisons are made, including those involving upright walking, reaction to colored picture book, differences in climbing ability, eating with a spoon and drinking from a glass, beginning of cooperative play, pointing to parts of the body, imitation of "writing," and affectionate behavior toward each other.

PCR-58-61: The Dynamics of an Experimental Neurosis: Its Development and Techniques for its Alleviation.

PCR-58: Conditioned Feeding Behavior and Induction of Experimental Neurosis in Cats (running time, 15.8 minutes). Cats are trained to respond to a light or bell signal by going to a food box into which food pellets are automatically released, and to obtain the food. An air blast blown just as the cat obtains the food is then employed to induce a motivational conflict. This induces inhibition of the feeding and a variety of "neurotic" patterns in and out of the experimental situation.

PCR-59: Effects of Environmental Frustrations and Intensification of Conflict in Neurotic Cats (running time, 12 minutes). Various types of environmental frustration are contrasted with those produced by the experimental motivational conflict.

PCR-60: Experimental Diminution of Neurotic Behavior in Cats (running time, 15 minutes). Four "therapeutic" techniques are demonstrated: (1) diminution of one of the conflicting drives by manual or forced feeding outside the cage; (2) retraining in the problem situation;

petting, gentle hand-feeding, "reassurance;" (3) environmental press - a maximally reinforced hunger drive and a movable barrier which slowly forces the animal closer to the food, resulting in a breakthrough of the cat's inhibitions and hurried gulping of the food; and (4) "social" example, set by a normal cat who has learned to feed at the signals - the neurotic cat gradually joins in the food-taking behavior.

PCR-61: Active Participation in Establishing More Satisfactory Adjustment (running time, 15 minutes). Normal cats are trained to depress a small disk platform which serves as a switch to activate their feeding signal. When the switch is turned off, or a barrier is placed between the cat and the food, the animals show various substitute responses; when the switch again works the signals, or the barriers are removed, the animals resume the normal feeding pattern. When these animals are shocked or given an air-blast at the moment of feeding, they develop all the neurotic behavior manifested by the cats in the previous reels, although generally in milder form. Most of the neurotic switch-workers, although at first avoiding the switch, gradually reexplore its use until they reestablish the self-signaling and feeding pattern, despite repetitions of the air blast.

These films were unfamiliar to almost all the students who participated in the experiment. In the case of the Cat Neurosis Series one student in each of two classes indicated that he had seen the four reels before; in the case of the Ape and Child Series between two per cent and six per cent of the students in each class had seen one or more of the reels.

The tests. Objective tests employing four-choice questions were constructed for each film series. The test for the Ape and Child Series included 78 items, 20 for each of the first two reels and 19 for each of the last two. The test for the Cat Neurosis Series included 80 items, 20 for each reel.[1]

---

[1] The tests and rating forms for all the films will be found in the microfilms for the dissertation, Appendix B.

Five scores were obtained for each test: number right on each of the four subtests, and total number right. All subjects were allowed and encouraged to answer all the items, and were instructed to guess when necessary. A check showed that less than one per cent of the items were omitted.

The reliabilities of these tests were estimated by the Kuder-Richardson method of rational equivalence (16). Use was made of the Kuder-Richardson formula number 20.[2] The reliability of the Ape and Child Series test, based on the scores for all the subjects in the experimental group, was .51. The reliability of the Cat Neurosis Series test was .73.

The interest rating form. An Interest Rating Form was devised to obtain the following data:

1. A roster of those who attended each film session;

2. An indication of the proportion of those who had already seen any of the films;

3. An indication of interest in the subject matter of the films; and

4. An indication as to whether or not the sessions were judged too short or too long, and as to whether or not there was any constant trend of interest (for the spaced methods groups).

Weights from zero to two or four (depending upon the number of choices permitted by the question) were assigned to the responses for each question except the first, which asked the subject whether he had seen the films before. The zero was assigned to the most negative response. The rating score was the sum of the weights for the responses for the seven questions scored. This score had a possible range from zero to 18. In addition, an analysis was made of the distribution of responses to each question separately.

Scheduling

The psychology classes used in this experiment met three times a week, on alternate days, for 50-minute periods. If the second period was held in the afternoon (on a Wednesday or Thursday), the first and third periods were held in the morning (on a Monday and Friday, or on a Tuesday and Saturday). If the second period was held in the morning, the first and third were held in the afternoon.

---

[2] The reliability coefficient $r = \frac{n-1}{n} \quad \frac{s^2 - \sum pq}{s^2}$, where n is the number of items, s^2 is the variance of the test scores, and $\sum pq$ is the sum of the item variances.

This class schedule made it impossible to conform exactly to the plan of having the same interval between the parts (in the four-part distributed presentation), since a weekend had to intervene. It also precluded the possibility of placing the test date one week from the mid-point of the series. Furthermore, during the course of the experiment it was found necessary to change a few dates in order to meet unexpected situations. The principal change involved delaying, for one class in each methods group for the Cat Neurosis Series, the test until approximately two weeks after the experimental showings.

It is believed that the "one-week" tests were administered far enough out on the retention curve to reduce to negligible proportions the effect of differences in the actual length of what was nominally a week, and that the same situation obtained with respect to the "two-week" retention tests.

Table 2 presents the schedule followed for the Ape and Child Series, and Table 3 presents the schedule for the Cat Neurosis Series. The time of day for each period is omitted; it has been pointed out that the periods for each class were staggered. The specific dates are also omitted. The Ape and Child Series was shown during February and March 1948, the Cat Neurosis Series was shown during the month of April 1948 to the same classes.

Procedures Followed

For the purposes of this experiment the films were used not as teaching aids but as total teaching instruments. The films were not described in any detail before presentation, they were not discussed in the classes during the course of the experiment, and they were not explicitly related to the rest of the content of the psychology course. This somewhat unusual procedure was followed in order to assure a maximum degree of uniformity with respect to the content presented in the films.

To achieve this uniformity, each film series was presented to each experimental class with a standard introduction which very briefly identified the series and explained in general the purpose of the study. The specific objective - that of comparing massed with spaced presentation - was not mentioned, so as not to prejudice rating of film length. The introduction was read at only the first session of the spaced presentations.

The massed presentation required a whole class period; the spaced presentations required only part of a period. However, in every case the first part of a spaced presentation was given at the beginning of the class period. In a few cases the succeeding parts were shown at the end of the period.

TABLE 2

EXPERIMENTAL FILM AND TEST SCHEDULE FOR APE AND CHILD SERIES

Method	Class Number	Day of Week											
		Mon	Tue	Wed	Thu	Fri	Sat	Mon	Tue	Wed	Thu	Fri	Sat
Control	4												Test
Control	9								Test				
Control	10											Test	
Control	11											Test	
1-Part	3				80-83*						Test		
1-Part	5		80-83						Test				
1-Part	8			80-83						Test			
2-Part	2			80-81		82-83						Test	
2-Part	7			80-81		82-83				Test			
4-Part	1	80		81		82		83				Test	
4-Part	6	80		81		82		83				Test	

---

* These are the PCR identification numbers for the reels of the Ape and Child Series as given in the Psychological cinema register catalogue (26). See page 17-18 of this report for the film titles.

TABLE 3

EXPERIMENTAL FILM AND TEST SCHEDULE FOR CAT NEUROSIS SERIES

		Day of Week														
Method	Class Number	Fri	Sat	Mon	Tue	Wed	Thu	Fri	Sat	Mon	Tue	Wed	Thu	Fri	Sat	Mon
Control	4										Test					
Control	9										Test					
Control	10															Test
Control	11															Test
1-Part	1	58-61*														Test
1-Part	2					58-61						Test				
1-Part	3				58-61						Test					
2-Part	6			58-59		60-61						Test				
2-Part	8					58-59		60-61								Test
4-Part	5				58		59		60		61				Test	
4-Part	7	58		59		60		61								Test

* These are the PCR identification numbers for the reels of the Cat Neurosis Series as given in the Psychological cinema register catalogue (26). See page 18-19 of this report for the film titles.

At the appointed time the experimenter, or an assistant, and a projectionist came to the scheduled class. Rating forms were distributed, and, for the first session, the introductory statement was presented verbatim.

In the second and succeeding sessions for the spaced presentation methods groups, an announcement was made merely to the effect that the second (or third, or fourth) film was to be shown.

At the end of each session, the rating forms were completed and collected by the experimenter. Thus, the one-part groups made one rating, on the entire series; the two-part groups made two ratings, one on each of the two pairs of two reels; and the four-part groups made four ratings, one on each of the four reels.

Projection facilities (screens, projectors, and an operator) were provided by the Audio-Visual Aids Library at the College. For the one-part and two-part showings two projectors were used to obviate the necessity of a time lag in setting up the second and succeeding reels.

On the whole, projection conditions were excellent, as most of the classrooms used were equipped with blackout curtains and wall screens.

Test scoring procedures. Test responses were recorded on IBM answer sheets which were later machine-scored.

The Experimental Population

The eleven classes of students taking the first (introductory) undergraduate course in psychology at the College which participated in the experiment included a total initial population of 545 students. However, after those subjects who missed one or more of the reels of either one series or the other, or failed to take the test for the series for which they saw all the reels, were eliminated, a sample of 460 was left for which complete data were available for the Ape and Child Series, and a sample of 410 was left for which complete data were available for the Cat Neurosis Series. Only 370 subjects provided complete data for both film series.

The principal analysis was conducted for each film series separately on the sample (of 460 or 410 subjects) for which complete data for that series were available. The sample of 370 subjects was used only to calculate the correlation between performance on the two film series tests.

Two indices were employed to determine whether the classes were equivalent in initial ability. These were the all-college grade-point average and the final course grade.

The all-college grade-point average. This is a weighted average of the grades earned by the student for his college work to date. The grades given at the College are "3" (highest), "2," "1," "0," "-1," and "-2." A grade of "-1" or "-2," is a failure.

Although some workers, such as Borow (2), have contended that the college grade-point average is a rather unsatisfactory measure of achievement or ability, it has been used in a very large number of studies as the principal, if not sole, criterion of college achievement, and has served to validate college entrance and college aptitude tests. Use has been made of it in this way at The Pennsylvania State College by Borow (2), Coblentz (7), Castore (5), Roulette (28), Schultz (29), Whittaker (36), and Mertens (24), among others.

Furthermore, several studies indicate a high degree of reliability in the sense of stability from year to year. Weaver (35) reported a reliability for the grade-point average over a four-year period of .88. Castore (5) found a correlation of .82 between the first semester grade-point average and the final grade-point average for all curricula at The Pennsylvania State College. Mertens (24) reported a correlation of .90 between third semester grade-point average and grade-point average to date for sophomore women in the education curriculum, and a correlation of .87 for sophomore women in the liberal arts curriculum.

This index may therefore be accepted as a relatively stable measure of general academic achievement at the College.

Final psychology grade. This variable cannot be considered a stable index of achievement, due to probable differences in instructor rating practices and requirements. However, it generally correlated to a greater extent with the film test scores than the all-college grade-point average, and was therefore retained as a matching variable.

Table 4 presents the means and standard deviations of the all-college grade-point averages and final psychology grades for the subjects for whom complete data were available for the Ape and Child Series; Table 5 presents the same information for the population for the Cat Neurosis Series. The following observations seem pertinent. First, the two indices are moderately correlated with each other (r = .5). Second, the all-college grade-point average is somewhat less variable than the final psychology grade in both cases: for the Ape and Child Series the grade-point average means for the classes ranged from 1.186 to 1.551, while the final psychology grade

TABLE 4

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL-COLLEGE GRADE-POINT AVERAGE (GPA) AND FINAL PSYCHOLOGY GRADE (PG), FOR SUBJECTS SEEING THE *APE AND CHILD SERIES*, GROUPED BY CLASS AND METHOD OF PRESENTATION

Method	Class	n	GPA		PG		r
			Mean	SD	Mean	SD	
Control	4	40	1.431	.695	1.175	1.022	.616
	9	43	1.368	.606	1.186	1.105	.357
	10	40	1.328	.637	1.225	.790	.605
	11	42	1.467	.608	1.238	.895	.557
Total		165	1.399	.639	1.206	.963	.519
1-Part	3	29	1.337	.692	1.283	.969	.572
	5	38	1.347	.583	1.158	.904	.548
	8	44	1.390	.554	1.591	1.007	.643
Total		111	1.361	.600	1.414	.982	.581
2-Part	2	41	1.350	.743	1.341	.873	.805
	7	60	1.186	.516	1.500	.806	.285
Total		101	1.253	.642	1.436	.837	.511
4-Part	1	43	1.551	.689	1.698	.977	.597
	6	40	1.423	.606	1.400	.800	.340
Total		83	1.489	.653	1.554	.909	.501
All Classes		460	1.374	.634	1.370	.941	.523

TABLE 5

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL-COLLEGE GRADE-POINT AVERAGE (GPA) AND FINAL PSYCHOLOGY GRADE (PG), FOR SUBJECTS SEEING THE CAT NEUROSIS SERIES, GROUPED BY CLASS AND METHOD OF PRESENTATION

Method	Class	n	GPA		PG		r
			Mean	SD	Mean	SD	
Control	4	37	1.482	.692	1.189	.982	.624
	9	41	1.378	.649	1.366	.904	.507
	10	34	1.322	.591	1.176	.706	.439
	11	47	1.388	.627	1.191	.891	.531
Total		159	1.393	.638	1.233	.884	.535
1-Part	1	47	1.524	.700	1.596	.938	.611
	2	41	1.334	.657	1.293	.862	.644
	3	21	1.440	.695	1.524	1.006	.570
Total		108	1.437	.689	1.268	.934	.620
2-Part	6	37	1.480	.600	1.205	.787	.386
	8	32	1.423	.514	1.656	1.049	.642
Total		69	1.453	.563	1.522	.926	.489
4-Part	5	23	1.521	.525	1.348	.698	.274
	7	50	1.235	.523	1.300	.849	.329
Total		73	1.325	.540	1.521	.813	.264
All Classes		410	1.403	.625	1.395	.902	.509

means ranged from 1.158 to 1.698; for the *Cat Neurosis Series* the means for the grade-point average ranged from 1.235 to 1.524, while the means for the final psychology grade ranged from 1.176 to 1.656. In both cases, the intra-class variabilities are greater for the final psychology grades than for the all-college grade-point averages. Third, attrition of the sample (as between the *Ape and Child Series* population and the *Cat Neurosis Series* population) resulted both in a constriction in the range of the means for both indices, and in a slight increase in the means. For example, the all-college grade-point average mean for the entire complete data group for the *Ape and Child Series* was 1.374 (Table 4), while the comparable mean for the *Cat Neurosis Series* was 1.403. These observations are consistent with the general hypothesis that better students attend class more regularly. It is believed, however, that while there may have been a slight amount of self-selective sampling, it was not sufficient to disturb the experiment seriously.

An analysis of variance was made for these two indices, for each film series separately.[3] There is no evidence that either the classes or methods groups are heterogeneous with respect to *all-college grade-point average*. The *classes within methods* are not significantly heterogeneous with respect to the *final psychology grade*, and for both samples there is but slight evidence of heterogeneity (F-ratio significant at the five per cent level of confidence) among the methods groups with respect to this variable. The conclusion may be drawn that the groups are essentially comparable in initial status.

---

[3] All the analysis of variance and covariance tables, showing sums of squares and crossproducts, and mean squares, have been omitted from this report. The tables are available on microfilm.

III. THE PSYCHOLOGY CLASSES' EXPERIMENT: RESULTS

The principal analysis for this experiment was based on the total test scores for each series. The means of the classes and the methods groups were compared by an analysis of variance and, for each film series, the effect of initial status[1] on test performance was determined by a covariance analysis that took into account final psychology grade and grade-point average.

An analysis was then made of the comparability of the methods as reflected in the ratings.

Finally, using only the data from those subjects who had seen both series and taken both tests, the correlation between achievement on the two tests was calculated.

The detailed results will be reported for each film series separately. Subject to minor variations, however, these results may be summarized for both film series as follows:

1. The experimental classes made very appreciable "gains" in comparison with the control classes who did not see the films. In other words, learning took place.

2. The differences among the three methods of presentation, as measured by the total test score, are consistently in favor of the spaced methods of presentation as opposed to the single massed session. However, these differences are negligible and insignificant, whether the test is administered at the end of one week or at the end of two weeks.

3. This finding also applies in general to the subtest scores.

4. There is a slight but significant relationship between test performance and initial status, but the lack of significant differences among the three methods cannot be attributed to differences in the initial status of the groups.

5. Analysis of the rating scores provides no reliable or consistent indication of discrimination among the three methods of presentation for the Ape and Child Series, and the ratings are independent of test performance. However, a small proportion of the subjects in the massed-method group reported that the film series was "too long." The ratings for the Cat Neurosis Series suggest the possibility of differentiating among the methods. In the massed presentation group between 60 and 80 per cent of the subjects rated the session "too long."

---

[1] While "final psychology grade" is not properly a measure of "initial status" it is sufficiently close to being so to permit use of the term "initial status" to denote both matching variables.

In short, although superficial examination of the raw scores might suggest some slight evidence favoring the spaced method of presentation, in the analysis of the data this difference fails to be statistically significant. The results from the psychology experiment indicate that, within the limits of sampling error, learning will be approximately as efficient if the learner is presented with a body of material in a single film session lasting one hour, as if he is presented with the same material in four 15-minute sessions spaced approximately equally over a week.

The Ape and Child Series

The basic data for the Ape and Child Series test - means, standard deviations, and standard errors for the total test score and the subtest scores - are presented in Table 6. The mean score for the massed (one-part) method group represents a gain of 22.75 points over the control group mean, while the mean scores for the two spaced methods groups represent gains of 23.51 and 23.44 points, respectively, over the control group. The films effected a significant amount of learning as measured by this test. Furthermore, while there is a suggestion of a consistent difference in favor of the spaced methods, this consistency is not sustained when the means for the classes are compared. Thus, while Classes 3 and 5 have lower (total) means than any other class, the mean of Class 8 exceeds the mean of one class in each of the two spaced methods groups. Furthermore, all three classes in the massed method group have higher means than any of the other classes on the second subtest (Part 2).

Comparison of the presentation methods. An analysis of variance was made for the total test score and for the subtest scores to determine whether any of the differences exceeded chance expectation. The results are reported in Table 7.[2]

It may be noted that the inter-class and inter-methods differences among the experimental groups on the total test score are statistically insignficant. The F-ratio for the methods is less than 1.[3]

The same thing is true for the first and third subtests. A slight lack of homogeneity is indicated for the second and fourth subtests. However, review of the means (Table 6) shows that for the second subtest the difference is in favor of the

---

[2] See footnote 3, Page 27.

[3] This may be interpreted as meaning that the variation between methods is less than the variation within the methods.

TABLE 6

MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR APE AND CHILD SERIES:

TOTAL TEST SCORE AND SUBTEST SCORES

			Total Test			Part 1			Part 2			Part 3			Part 4		
Method	Class	n	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm
Control	4	40	31.32	5.44	.87	9.56	3.14	.50	7.10	3.04	.49	5.77	1.80	.29	8.15	2.35	.38
	9	43	28.88	4.05	.63	9.65	2.60	.40	6.86	1.65	.25	5.28	1.77	.27	7.05	2.23	.34
	10	40	33.47	5.05	.81	11.75	2.37	.38	7.52	3.31	.53	5.45	1.56	.25	8.17	2.13	.34
	11	42	32.67	4.99	.78	11.33	2.59	.40	7.71	2.05	.32	5.38	1.90	.30	8.07	2.05	.32
Total		165	31.55	5.20	.41	10.56	2.86	.22	7.35	2.21	.17	5.47	1.77	.14	7.85	2.24	.17
1-Part	3	29	53.97	5.22	.99	14.55	1.92	.36	14.66	2.49	.47	11.62	2.07	.39	13.45	1.97	.37
	5	38	53.47	4.89	.80	13.97	2.03	.33	14.08	2.00	.33	12.16	1.69	.28	13.26	1.96	.32
	8	44	55.23	3.81	.58	14.23	1.79	.27	14.91	1.93	.29	11.91	1.63	.25	13.93	1.60	.24
Total		111	54.30	4.66	.44	14.22	1.92	.18	14.56	2.14	.20	11.92	1.79	.17	13.58	1.85	.18
2-Part	2	41	55.37	4.25	.67	14.76	3.36	.53	13.46	3.27	.52	12.10	2.99	.47	13.95	2.91	.46
	7	60	54.87	4.56	.59	15.02	4.59	.59	13.92	2.00	.26	12.07	1.89	.25	14.22	2.05	.27
Total		101	55.06	4.44	.44	14.91	4.14	.41	13.73	2.60	.26	12.08	2.40	.24	14.11	2.44	.24
4-Part	1	43	54.30	5.11	.79	15.16	2.74	.42	13.88	2.69	.42	11.26	2.18	.34	14.70	1.72	.26
	6	40	55.72	4.12	.66	15.07	1.46	.23	13.63	2.40	.38	12.25	2.19	.35	14.82	1.39	.22
Total		83	54.99	4.71	.52	15.12	2.22	.24	13.76	2.56	.28	11.73	2.24	.25	14.76	1.57	.17
Total Experimental Group		295	54.76	4.62	.27	14.71	2.96	.17	14.05	2.46	.14	11.92	2.15	.12	14.09	2.06	.12

TABLE 7

F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS:

APE AND CHILD SERIES TEST

	Total Test Score	Part 1 Score	Part 2 Score	Part 3 Score	Part 4 Score
Between experimental and control groups	2428.010***	242.673***	840.548***	1075.228***	905.623***
Between classes within experimental methods	1.351	< 1	< 1	1.269	< 1
Between experimental methods	< 1	2.502	3.849*	< 1	8.135***

* Significant at the 5% level of confidence.
*** Significant at the 0.1% level of confidence.

massed presentation. For the fourth subtest it is in favor of the two spaced methods. These differences are, in both cases, of the order of a score point, and cannot be considered of practical importance.

In every case, the difference between the experimental and control group is significant beyond the 0.1 per cent level of confidence.

The variance estimates for the total score were adjusted by means of a covariance analysis for "initial status" as measured by grade-point average to date and final psychology grade. The multiple correlation between these variables considered jointly and the total test score for the entire experimental group was .29.

Since the regression accounts for only about eight per cent of the total test score variance, it would require exceptionally large differences in initial status to change the relative standing of the groups with respect to the film test performance.

When the variance estimates for the total test score were adjusted for this multiple correlation, for both the experimental classes and methods groups there was a slight reduction in the estimate of error, and a slight increase in the mean square attributable to the "between groups." However, the F-ratios remain well below the five per cent level of significance. In short, the lack of differences among the experimental methods cannot be considered an artifact of initial differences among the groups, at least with respect to these two matching variables.[4]

The interest ratings. The *Ape and Child Series* is one of the most popular with students in psychology classes at The Pennsylvania State College. Kellogg's subjects, the child Donald and the chimpanzee Gua, are sprightly and charming, and the content of the films is readily grasped (if not completely remembered) by almost all students.

Analysis of the interest ratings merely adds statistical evidence to these observations. The highest possible interest score was 18 points. The mean for no session was lower than 13.7, and the ratings were very markedly massed at the high end of the scale. No subject had a rating score of less than 7.

---

[4] See footnote 3, Page 27.

TABLE 8

MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: APE AND CHILD SERIES

Method	Film Period	Class	n	Mean Rating	SD	r
1-Part	First (Only)	3	29	15.52	1.25	.024
		5	38	15.26	1.46	.019
		8	44	14.82	2.52	-.028
		Total	111	15.15	1.94	.005
2-Part	First	2	41	14.41	1.23	.029
		7	60	15.43	1.79	-.146
		Total	101	15.16	1.59	-.101
	Second	2	41	14.76	1.22	.071
		7	60	15.43	1.79	-.146
		Total	101	15.16	1.59	-.101
4-Part	First	1	43	14.35	1.31	-.003
		6	40	14.47	1.84	-.157
		Total	83	14.41	1.59	-.071
	Second	1	43	13.91	1.07	-.026
		6	40	14.55	1.56	-.031
		Total	83	14.22	1.37	-.071
	Third	1	43	13.69	1.30	.134
		6	40	14.50	1.40	.135
		Total	83	14.08	1.41	.190
	Fourth	1	43	13.91	1.27	-.151
		6	40	14.92	1.38	.136
		Total	83	14.40	1.42	-.006

Since a rating was made at the end of each session, the massed method classes made one rating, based on all four reels. The two-part method classes made two ratings, the first on reels one and two, the second on reels three and four. The four-part method classes made four ratings, one after each reel.

The initial hypothesis was that the mean rating of the massed presentation group would be lower than that for any other session, on the grounds that exposure to an hour-long film would bring more than satiation and would yield negative responses to such questions as "Would you like to see more films in this series?" or "Did this film hold your attention?" With respect to the spaced methods groups, it was thought that rating means would be higher in the first session than in the later sessions.

This hypothesis was not sustained. Table 8 presents the mean rating scores by session, class, and methods group, and the correlations between the ratings and the test scores for the subtest or tests for the reels on which the rating was based.

Although some of the inter-sessions differences may be significant, they were considered too small to merit statistical analysis. The only observation with respect to these mean scores which seems of any importance is that the means for the classes participating in the one-part (massed) presentation were slightly higher than the means for almost all the other classes.

None of the correlations between the ratings and the test scores differ significantly (at the five per cent level of confidence) from zero. In other words, test performance was essentially independent of attitude toward the film series or the presentation method, as reflected in these ratings. The considerable skewing and restriction in range of rating scores may have contributed to reducing the correlations. However, as will be pointed out below, even for those series for which the ratings were more widely distributed, the correlations between test performance and rating score were of about the same order as reported here.

An analysis was also made of the per cent distribution of responses to the eight questions included on the rating form. This analysis revealed the popularity of this film series to no small extent. Except for one student in a single session in each of the classes exposed to the four-part presentation method, everyone reported that the films held their interest most or all of the time. For all classes and sessions but two, the film or films shown were rated by a majority as "very good" or "excellent."

In point of fact, of the eight questions asked, only one, that directly relating to film length, provides any consistent discrimination among the methods groups. No one in the two spaced methods groups thought the sessions (lasting either 15 or 30 minutes) too long, while 3.4 per cent, 4.5 per cent, and 10.5 per cent, respectively, of the three classes exposed to the massed presentation reported that they thought the film was too long. On the other hand, between 10.5 per cent and 24.2 per cent of the massed presentation classes thought the session too short.

While a small proportion of the subjects thought that the hour session was too long, this length has not been demonstrated to have had a deleterious effect either on reported interest or measured learning.

The Cat Neurosis Series

The general conclusions to be drawn from the analysis of the test scores and rating scores for the Cat Neurosis Series of films are substantially the same as those reported above for the Ape and Child Series. The basic test data for the Cat Neurosis Series are presented in Table 9. It will be remembered that, for this series, one class in each methods group took the test two weeks after the experimental showings, while the remaining classes took the test one week after the showings. This difference had a significant effect on test performance. The mean total score of the subjects in the one-week retention group represents a gain of 16.68 points over the mean of the control group, while the mean score for the two-week retention group represents a gain of 13.65 points. Furthermore, the mean total score for each method in the one-week group is higher than the mean score for any method in the two-week retention group. In other words, some forgetting took place.

For both the total test score means and the subtest score means within each retention group there is a consistent trend favoring the spaced presentation methods. These differences, however, are consistently small.

Comparison of the presentation methods. An analysis of variance was made for both the total scores and the subtest scores to determine whether (1) the inter-methods differences were significant, (2) the inter-retention period differences were significant, (3) there was any inter-action between the retention period and the presentation method. Significant interaction could be interpreted to mean that the relative effectiveness of the methods, as measured, depended in part upon the length of time that elapsed between the experimental sessions and the test. For example, one might find that

TABLE 9

MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR CAT NEUROSIS SERIES:

TOTAL TEST SCORE AND SUBTEST SCORES

			Total Test			Part 1			Part 2			Part 3			Part 4		
Method	Class	n	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm
Control	4	37	22.68	4.41	.73	4.89	2.13	.36	6.65	2.26	.38	5.11	1.91	.32	6.03	2.05	.34
	9	41	22.51	4.52	.71	5.02	1.55	.25	6.10	2.14	.34	5.56	1.85	.29	6.05	1.99	.31
	10	34	22.50	3.65	.63	4.94	1.45	.25	5.88	1.89	.33	5.53	1.83	.32	6.15	1.73	.30
	11	47	22.36	4.09	.60	5.57	1.55	.23	6.21	2.00	.30	5.13	1.89	.28	5.53	1.85	.27
Total		159	22.50	4.19	.33	5.14	1.71	.14	6.21	2.09	.17	5.32	1.88	.15	5.91	1.93	.15
One Week Retention Group																	
1-Part	2	41	38.24	9.38	1.48	9.90	2.83	.45	9.83	3.14	.50	10.10	2.91	.46	8.63	3.05	.48
	3	21	39.19	6.31	1.41	10.05	2.13	.47	9.19	2.67	.60	11.05	2.63	.59	8.76	2.09	.47
Total		62	38.56	8.48	1.09	9.95	2.61	.33	9.61	3.00	.38	10.42	2.85	.36	8.68	2.76	.35
2-Part	6	37	39.27	7.85	1.31	10.22	3.00	.50	9.84	2.78	.46	10.65	2.42	.40	8.78	2.62	.44
4-Part	5	23	40.70	6.86	1.46	10.83	1.81	.39	10.17	2.30	.49	10.87	3.71	.79	8.83	1.71	.36
Total One Week Group		122	39.18	8.04	.73	10.20	2.63	.24	9.79	2.82	.26	10.57	2.92	.27	8.74	2.55	.23
Two Week Retention Group																	
1-Part	1	47	34.23	8.21	1.21	8.32	2.13	.31	8.90	2.70	.40	9.19	2.66	.39	7.74	2.98	.44
2-Part	8	32	37.12	8.30	1.49	9.53	2.03	.36	10.44	2.81	.50	9.37	3.04	.55	7.75	2.69	.48
4-Part	7	50	37.34	4.84	.69	10.18	2.10	.30	9.72	1.95	.28	9.64	2.05	.29	7.80	2.37	.34
Total Two Week Group		129	36.15	7.27	.64	9.34	2.25	.20	9.63	2.53	.22	9.41	2.55	.23	7.77	2.68	.24
Total Experimental Group		251	37.62	7.80	.49	9.76	2.48	.16	9.71	2.68	.17	9.98	2.80	.18	8.24	2.67	.17

TABLE 10

F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS:

CAT NEUROSIS SERIES TEST

	Total Test Score	Part 1 Score	Part 2 Score	Part 3 Score	Part 4 Score
Between experimental and control groups	502.202***	422.078***	194.034***	339.861***	90.577***
Between methods: one week group	< 1	< 1	< 1	< 1	< 1
Between methods: two weeks group	2.633	9.484***	3.307*	< 1	< 1
Between retention periods	12.779***	13.169***	< 1	12.671***	8.730***
Interaction: retention periods and methods	< 1	1.220	1.458	< 1	< 1

* Significant at the 5% level of confidence.
*** Significant at the 0.1% level of confidence.

although there was some loss in retention as between the one week group and the two week group, the subjects seeing the films in the massed presentation forgot a relatively greater amount than subjects seeing the films in either of the spaced presentations.

The results of the analysis of variance are presented in Table 10.[5]

The following conclusions are justifiable:

1. The difference between the experimental and control groups is highly significant (for both the total test score and the subtest scores the F-ratio exceeds that required for significance at the 0.1 per cent level of confidence).

2. With respect to the one-week retention group, there is no evidence of significant differences among the three presentation methods.

3. With respect to the two-week group, there is evidence that the means for the three methods do differ significantly on the first subtest (the one-part method mean was 8.32, as compared to the four-part method mean of 10.18). This difference is significant at the 0.1 per cent level of confidence. For the subtest on the second reel the differences are smaller and significant only at the 5.0 per cent level of confidence. These findings, however, are based upon data which are too meager to support any general conclusion with respect to the underlying principle involved. One is tempted to conjecture that retroactive inhibition was relatively more potent with respect to the earlier rather than the later reels, and that its effects were increasingly marked as the retention period was lengthened. The controls used (or not used) in this study, however, provide no basis of support for this conclusion.

The inter-method differences for the total test score and the remaining two subtests scores for the two-week retention group are not significant.

4. The differences between the one-week and two-week retention groups are highly significant (at the 0.1 per cent level of confidence) for all the scores except those for the subtest on the second reel. Examination of the data reveals, for this latter subtest, greater intra-method than inter-method differences.

---

[5] For this analysis, Classes 2 and 3 (one-part method, one-week group) were combined, to simplify the calculations. This combination was justifiable, since the means for the two classes did not differ significantly.

5. There is no evidence that a significant interaction occurred between the length of the retention period and the method of presentation. That is, within the limits of experimental error, the inter-method differences for the two-week retention group were proportional to the inter-method differences for the one-week retention group.

The variance estimates for the total score were adjusted by means of a covariance analysis for "initial status" as measured by grade-point average to date and final psychology grade.

For the one-week group, the muliple correlation between the total test score, on one hand, and the grade-point average and final psychology grade, on the other, was .38; for the two-week group the multiple correlation was .39. Adjusting the variance estimates led to a slight increase in precision (as indicated by the reduction in the error mean square), but the adjustments do not change the pattern. The lack of significant differences among the methods means cannot be attributed to inequality in initial status of the groups as measured by the all-college grade-point average to date and the final psychology grade.[6]

The interest ratings. The rating score means and correlations with the test scores are given in Table 11. It may be noted that here, as with the Ape and Child Series, the amount learned is practically uncorrelated with interest as measured.

The Cat Neurosis Series, however, is not nearly as popular with first-year psychology students as is the Ape and Child Series. The Cat Neurosis Series is more static, the concepts it tries to convey are more complex, and the subjects (cats in boxes) not as interesting or appealing. The mean ratings reflect this to a slight extent. Furthermore, it is probable that introductory psychology students lack the background necessary to comprehend these films. The films are perhaps too technical.

The distribution of ratings is skewed toward the low end of scale, and the mean ratings are slightly lower than those for the Ape and Child Series.

Furthermore, the mean rating score for the massed presentation sessions is between one and two points lower than the mean rating score for any other session, indicating a possible differentiation among the methods.

---

[6] See footnote 3, Page 27.

TABLE 11

MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: CAT NEUROSIS SERIES

Method	Film Period	Class	n	Mean Rating	SD	r
1-Part	First (Only)	1	47	9.81	2.51	.033
		2	41	11.76	2.29	.131
		3	21	12.43	1.36	.377
		Total	109	11.05	2.50	.202
2-Part	First	6	37	12.76	2.90	.541**
		8	32	13.59	2.71	.274
		Total	69	13.14	2.84	.426**
	Second	6	37	12.62	2.71	.261
		8	32	12.72	3.11	.257
		Total	69	12.67	2.91	.246
4-Part	First	5	23	13.52	1.25	-.114
		7	50	13.68	1.92	.064
		Total	73	13.63	1.74	.021
	Second	5	23	12.52	1.77	-.076
		7	50	12.00	2.13	.140
		Total	73	12.16	2.03	.085
	Third	5	23	13.13	1.80	.445*
		7	50	13.42	2.24	-.019
		Total	73	13.33	2.11	.127
	Fourth	5	23	13.13	1.54	.176
		7	50	13.52	2.00	.106
		Total	73	13.40	1.88	.096

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

Analysis of the distribution of responses to the questions on the rating form indicates some of the factors underlying this difference. The most important one was the length of the session: between 61 per cent and 81 per cent of the students subjected to the single massed session thought the session too long; between 13 per cent and 21 per cent thought the two 30-minute sessions, too long. Furthermore, slightly higher percentages of the students in the massed method classes reported that the series held their interest only "some of the time" or "none of the time," and higher percentages in two of these classes rated the films "useless" than was the case for classes participating in either of the two spaced methods groups.

It is interesting to note, however, that the most critical class (Class 1) was also the most critical class with respect to the Ape and Child Series. In that series, Class 1 was used in the four-part method group, and a higher proportion of students in that class rated each reel "fair" or "poor" than in any other class. The generally negative reaction of Class 1 to the film series therefore seems, in part at least, to be a function of the attitude of the class to films in general.

One may conclude on the basis of the evidence presented above that the judgment as to whether a film session is too long is a function, to a significant extent, of the particular film series and the preparation of the subjects for it. In the one case, at most ten per cent of any class thought the hour session was too long. In the other case, 80 per cent in two classes, and 60 per cent in the third, thought the hour session was too long; and sizeable fractions of each class thought the 30-minute sessions too long.

However, the greater length of the massed session, even if judged negatively, has not been demonstrated, for the Cat Neurosis Series or for the Ape and Child Series, to have a deleterious effect on learning. The data do suggest that it would be preferable to produce and use interesting films for long sessions.

Relationship between Test Performances on the Two Film Series

The same classes participated in the experiment with each psychology film series. Except for Class 3 (and the four control classes), no class saw both film series in the same method of presentation.

When those subjects were eliminated who failed to take both tests (in the control group), or who failed to see all four reels of both series and take both tests (in the experimental group),

TABLE 12

MEANS FOR, AND CORRELATIONS BETWEEN, APE AND CHILD SERIES AND CAT NEUROSIS SERIES TOTAL TEST SCORES: COMPLETE DATA CASES ONLY

Class	No. of Cases	Ape and Child Method	Ape and Child Mean	Cat Neurosis Method	Cat Neurosis Mean	r
4	36	Control	31.03	Control	22.61	-.13
9	38	Control	28.79	Control	22.71	-.18
10	34	Control	33.00	Control	22.50	.04
11	41	Control	32.63	Control	23.05	.20
Total Control Group	149		31.46		22.73	-.02
1	38	4-Part	54.53	1-Part	35.13	.11
2	34	2-Part	55.65	1-Part	38.18	.36*
3	19	1-Part	54.58	1-Part	38.74	.13
5	21	1-Part	54.90	4-Part	40.19	.46*
6	32	4-Part	56.03	2-Part	39.22	-.01
7	48	2-Part	55.06	4-Part	37.37	.34*
8	29	1-Part	55.59	2-Part	37.34	-.13
Total Experimental Group	221		55.21		37.76	.14

* Significant at the 5% level of confidence.

the total sample decreased to 370, of whom 149 were in the control classes and 221 in the experimental classes.

Table 12 presents the means for both film tests and the correlations between the two film tests for this reduced "complete data" sample. No analysis of variance was made of the methods differences from this data; inspection of the means suggests that such an analysis would not change the conclusions already arrived at. It is interesting to note that attrition of the sample led, for almost every class, to a slight increase in the means. This is consistent with the observation that the better students tend to have better attendance records.

The correlations between the two tests are somewhat more interesting. With three negligible exceptions, these correlations do not differ significantly from zero.

There may be several possible reasons for this result. First, the constriction in range implied above would tend to reduce the correlations. Second, the low reliabilities of the tests would contribute to low inter-test correlations. Third, it may be that the contents of the two films were so dissimilar that the tests measured very different things. Fourth, it may be that the three methods were differentially effective with different learners so that a student who might learn efficiently from a massed presentation would not learn efficiently from the spaced presentation, while another would learn equally well from both, and a third would learn best from the spaced presentation. However, aside from the relatively low reliability indices for these tests, there is no information available to decide the issue.

IV. THE NAVY EXPERIMENT: DESIGN AND PROCEDURES

The Navy experiment was conducted with apprentice seamen at the Great Lakes Naval Training Center, Great Lakes, Illinois. The plan for this replication called for the use of ten companies of apprentice seamen with two series (each comprising three reels) of films - a series on Elementary Hydraulics and a series on Rules of the Nautical Road.

Two companies were designated as control groups. Each control group was administered one of the two tests, to provide a base line against which to measure learning. Early in the negotiations looking toward arranging the Navy replication it was made clear that two hours was the maximum amount of time that could be permitted for any one recruit. Therefore, each experimental group was shown only one of the film series, in one spacing pattern. Two companies were each shown the series on Elementary Hydraulics in one 45-minute period; two were shown the Hydraulics Film Series in three 15-minute periods. The Rules of the Nautical Road Series was shown in the same manner. For each experimental group the test was scheduled approximately one week from the mid-point of the presentation schedule. The program began on Tuesday, 10 August 1948. Four and one-half days were devoted to the showings and the following week to the testing.[1]

Films and Tests

The criteria[2] imposed in the selection of the film series of the psychology classes' experiment were used also for the selection of film series for the Navy replication.

The films. The series used included the following:[3]

Elementary Hydraulics Series MN-1730.

MN-1730d - Application of Pascal's Law (running time 12 minutes, sound, black and white). Pascal's Law is demonstrated in animation, and the input and output pistons and the working of a simple schematic hydraulic system are presented. The way in which work is done by means of a hydraulic system is shown, as well as measurement of work, including the basic formulas for calculating work done and force applied.

MN-1730e - Liquids in Motion (running time, 13 minutes, sound, black and white). The principles of liquids at rest are reviewed and pressure energy and velocity energy are defined and illustrated. The relation of energy output to resistance to flow is explained, and variation in energy output with three different lengths of pipe is illustrated.

MN-1730g - Simple Hydraulic Systems (running time, 17 minutes, sound, black and white). The hydraulic jack and the hydraulic braking system of an automobile are used to demonstrate the features of simple hydraulic systems.

---

[1] See the microfilm of the original dissertation, Chapter V, for a detailed explanation of the reasons for limiting each series to three, rather than employing four, reels.

[2] See Page 16.

[3] Descriptions based on entries in the Catalog of training films for the United States Navy and Marine Corps (32).

Rules of the Nautical Road Series MN-202.

MN-202f - Lights of Vessels Being Towed (running time, 11 minutes, sound, color). Light specifications for vessels being towed are given, including lights for barges, canal boats, scows, and dump scows. Covers inland waters and international rules.

MN-202i - Visual Day Signals (running time, 14 minutes, sound, color). Identifying day signals for indicating the vessel's occupation (fishing vessel, cable ship, dredge, etc.) are shown, as well as international warning and distress signals.

MN-202j - Whistle Signals for Approaching Steam Vessels (running time, 17 minutes, sound, black and white). The rules for using the one-, two-, and three-blast signals, the danger signal, and the bend signal are illustrated in several examples of approach situations.

All the films described above used animation throughout. The Elementary Hydraulics Series presented essentially a rational learning task in which the subject was taught the principles relating a very small number of variables in model hydraulic systems. The Rules of the Nautical Road Series presented a rote memory task, in which the subject was expected to learn a set of arbitrary rules governing basic signalling at sea.

These films were unfamiliar to almost all the men in the experiment. In the case of the Elementary Hydraulics Series, from 1.1 to 4.0 per cent of each company had seen one or more of the reels. In the case of the Rules of the Nautical Road Series, 2.5 per cent of the men in each of two companies had seen the third reel of the three reels making up the series. No subject had seen the first two reels.

The tests. Objective tests employing four-choice questions were constructed for each series. Each test included 15 items for each reel in the series covered. To facilitate scoring, the questions for the reels were kept separate. Since each series originally included four films (one in addition to the three described above) each test originally had 60 items. When it became evident that one reel would have to be eliminated, it was found feasible to eliminate, in each case, the film covered by the last 15 items of each test. The subjects answered all the questions, but the analysis was made of the score on the first 45 items only. The score in each case was total number right. The subjects were instructed to answer all questions, guessing if necessary. A check of the test sheets indicated that less than one per cent of the questions were unanswered. Both total test and sub-test scores were recorded. The reliabilities of these tests also were estimated by the Kuder-Richardson method of rational equivalence. On this basis, the reliability of the test on the Elementary Hydrualics Series, based on the entire experimental group, was .80; and the comparable reliability for the Rules of the Nautical Road Series was .57.

The rating form. An Interest Rating Form was devised to collect the following data:

1. A roster of those who attended each film session;
2. An indication of the proportion of each group which had already seen any of the films;
3. An indication of interest in the subject matter; and
4. An indication of the judged adequacy of visibility and acoustics.

Weights from zero to two or four (depending upon the number of choices permitted by the question) were assigned to the responses for each question except the first, which asked the subject whether he had seen the film before, and the last two, dealing with the physical surroundings. The zero was assigned to the most negative response. The rating score was the sum of the weights for the responses for the four questions scored. This score had a possible range from 0 to 6. In addition, an analysis was made of the distribution of responses to each question separately.

Scheduling

An attempt was made in the planning stage to distribute the film presentations so that effects of miscellaneous variables such as the time of day and previous activity would be randomized. However, in discussions at the Station it was emphasized that the long single period should be placed at the end of the day, since the time for all other periods was strictly limited to 50 minutes.

Station problems - "must" classes, previous commitments, and similar factors - made it impossible, furthermore, to achieve a randomized distribution of the other experimental sessions. A series of changes in the "ideal" plan for the spacing pattern were therefore made to meet these problems.

The final schedule, as actually followed, is given in Table 13. As will be seen, the spacing pattern was not consistent for the experimental groups. Furthermore, the tests

TABLE 13

FILM PRESENTATION AND TESTING SCHEDULE FOR NAVY REPLICATION

Period	Tues	Wed	Thurs	Fri	Sat	Mon	Tues	Wed	Thurs	Fri
1	210/R-f	210/R-j	211/H-e	210/R-i			202/TR	208/TH	204/TH	205/TR
2	206/H-d	211/H-d	212/R-i			212/R-j				
3		212/R-f	206/H-g					209/TR	206/TH	
4	206/H-e						210/TR	211/TH	212/TR	
5			211/H-g							
6	202/R-all	204/H-all	203/H-all	205/R-all				203/TH		

Note: In each cell the company number, film series, and particular film are indicated. The capital letter R stands for the Rules Series; the capital letter H stands for the Hydraulics Series. The lower case letter stands for the particular film in the series. The cells for the second week indicate the test dates - thus TR stands for the test on the Rules Series, TH for the test on the Hydraulics Series.

did not always fall one week from the mid-point of the film presentation. The interval varied from six to eight days.

It is to be noted that the intervals are not comparable except in the most generous view, and that the "time of day" variable was confused beyond any hope of unscrambling. However, since the results suggest that spaced presentation is neither better nor worse than massed presentation, the specific spacing employed does not seem crucial.

Procedures Followed

After a firm schedule had been arranged with the Scheduling Officer, notices were sent out to the companies assigned to participate in the experiment advising them of the date, time, and place of the first session they were to attend. Those companies exposed to the three-part presentation method did not know the dates of the second and third part showings until the day before in each case.

At the appointed time each company appeared in march formation. The company entered the projection room in double file. Pencils and the rating forms were handed out at the door by two proctors at the beginning of each showing, and collected at the end. A rating was collected at the end of each showing. Thus, the one-part group made one rating and the three-part groups three ratings.

The film presentation. An announcement introducing the film series was presented substantially verbatim to each experimental company at its first session. The men were told that they were participating in an experiment dealing with training films, and that they were to rate the films and be tested after the films were shown.

Two "Ampro" sound projectors, and one speaker, were used in the massed presentations. The first reel was set up on one projector and the second on the other. While the second reel was being projected, the third was set up on the first projector.

Test procedures. The testing of the groups was carried out exactly according to schedule. Four proctors were assigned to the experiment from a service company. For testing, the chairs were moved into two rooms on the ground floor of the other wing of the building in which the experiment was conducted. Approximately 70 chairs were put into each room, spaced for maximum separation. Two proctors were assigned to each test room. Before the test they were responsible for placing an IBM pencil, an IBM answer sheet, and a test booklet on each chair. The experimenter in the room read a set of standard directions to the group. In addition to supplying identification information, the men were instructed to indicate, on the IBM answer sheet, whether they had previous training in elementary hydraulics or in the rules of the nautical road.

The Experimental Population

The ten companies of recruits comprised a total initial population of 1238 men. However, as a result of absences at one or more sessions, complete data are available on only 887 men. The companies, as reduced, ranged in size from 59 to 109 men. The oldest company (Company 202) had been formed approximately eight weeks before the experiment began. The youngest (Company 212) had been formed seven weeks before the experiment began. The men in these companies were all volunteers, having enlisted for a three-year term.

Comparability in initial ability. Five criteria were available to determine the initial comparability of the companies: age, previous education, Navy General Classification Test score, Mechanical Aptitude Test score, and previous knowledge of the subject matter of the film series used.

A preliminary analysis suggested that age and previous education were relatively unimportant variables. The mean age of the sample was 18.30 years, the mean educational status 10.88 years.. Over 99 per cent of the men in the sample were between 17 and 21 years old.

The third and fourth variables, Navy General Classification Test score and Mechanical Aptitude Test score, were retained as the principal matching variables, and employed in a covariance analysis to determine whether the differences (or lack of them) in the experimental test performance were significantly affected by differences in "general intelligence" and "mechanical aptitude". Table 14 presents the means and standard deviations for the General Classification Test and for the Mechanical Aptitude Test for the companies individually, for the methods groups, and for the total experimental group; figures are shown for each film series population. It will be noted that all these means and standard deviations closely approximate the population norms for each test (mean of 50, standard deviation of 10). Furthermore, the two tests are moderately highly correlated with each other.

The analysis of variance for each test permits the interpretation that the differences among the companies in initial status, as measured by the General Classification Test and the Mechanical Aptitude Test, may be attributed to chance fluctuations.[4]

---

[4] See footnote 3, Page 27.

TABLE 14

MEANS AND STANDARD DEVIATIONS FOR NAVY GENERAL CLASSIFICATION TEST AND MECHANICAL APTITUDE TEST, AND CORRELATIONS BETWEEN THE TESTS: FOR COMPANIES GROUPED BY FILM SERIES AND METHOD OF PRESENTATION

				GCT		MAT		
Film Series	Method	Company	n	Mean	SD	Mean	SD	r
Hydraulics	Control (No film)	208	109	52.79 53.03	8.76	49.51	9.23	.421
	1-Part	203	114	50.82	11.62	48.46	10.49	.676
		204	76	52.21	9.29	49.25	9.76	.539
	Total		190	51.38	10.77	48.78	10.21	.630
	3-Part	206	59	51.81	9.46	50.09	8.82	.626
		211	93	50.12	9.33	48.67	7.50	.406
	Total		152	50.78	9.42	49.22	8.07	.503
	Total Experimental Group		342	51.11	10.20	48.97	9.32	.583
Rules of the Road	Control (No film)	209	84	50.50	11.03	48.92	10.67	.718
	1-Part	202	109	51.08	10.44	48.78	10.37	.627
		205	84	52.02	10.62	51.00	10.28	.614
	Total		193	51.49	10.53	49.75	10.39	.622
	3-Part	210	80	50.31	9.54	49.83	8.46	.580
		212	79	48.77	10.66	46.56	8.79	.556
	Total		159	49.55	10.14	48.20	8.78	.569
	Total Experimental Group		352	50.61	10.40	49.05	9.73	.602
	All Companies		887	51.06	10.22	49.06	9.61	.589

For the fifth variable, previous knowledge of the subject, the following criteria were used:

1. If a subject, to whom the *Elementary Hydraulics Series* was shown, checked on his answer sheet that he had had a course in physics in high school or elsewhere, he was counted as having previous knowledge of the content. Approximately one-third of each company indicated that they had had at least one such course.

2. If a subject, to whom the *Rules of the Nautical Series* was shown, checked on his answer sheet that he had read the section on "Rules of the Road" and/or the section on "Signals" in the *Bluejackets' manual* (31)[5], he was counted as having previous knowledge of the content of this series of films. In one company about a fifth of the men *claimed* to have studied these chapters. In the other four companies used with this film series not more than five or six per cent of the men had read the chapters in question.

It may be noted here that although the absolute learning gains differ depending on whether this variable is taken into account or not, the relative status of the presentation methods is unaffected. The distribution with respect to previous knowledge as defined here is considered in the next chapter where the effect of the variable on the test scores is discussed.

Since so few of the men had seen any of the films, this factor was not included in the analysis of the effects of previous knowledge.

V. THE NAVY EXPERIMENT: RESULTS

The results of the Navy experiment are, in general, similar to those of the psychology classes' experiment. Comparison between the experimental and control groups assures that a measureable amount of learning took place. Comparisons among the experimental companies and methods groups fail to provide evidence of a statistically significant difference in favor of either the massed (45-minute) single session or the three spaced (15-minute) sessions. These results remain unchanged when the film test scores are adjusted for initial status as measured by a combination of the Navy General Classification Test and the Mechanical Aptitude Test.

---

[5] Chapter 36, "Rules of the Road," pp. 429-436, covers the basic facts (but not all the details) of night lights and whistle signals. Chapter 46, "Pyrotechnics, Distress Signals and Storm Warnings," pp. 551-561, covers visual day signals.

The detailed results for each film series - the series on Rules of the Nautical Road and the Elementary Hydraulics Series will be discussed separately.

Rules of the Nautical Road Series

The basic data, including means and standard deviations for the companies and methods groups, for the total test score and the three subtest scores, are presented in Table 15. Inspection of the table suggests, first, that either the film series taught little or the test was a poor measuring instrument. The mean score for the best company (Company 202) was only about a third of the maximum possible on the test (45 points). The difference between the experimental and control group was only 3.3 score points.

Second, with the exception of Company 202, which was shown the series in one long session, the means for the experimental companies are practically identical. If the performance of Company 202 alone were used to represent the effect of the massed presentation, this method would demonstrate a statistically significant superiority. The data analysis suggests, as will be pointed out below, that this result may have been largely due to the initial superiority of the company with respect to the matching variables.

Preliminary to an analysis of the methods differences per se, the effect of previous knowledge on the test performance was studied.

Previous knowledge. Table 16 presents the analysis of the effect of previous knowledge on the test scores. There is, for each methods group, and for each company with the exception of Company 202, a consistent difference in favor of those subjects in each company who reported having studied the "Rules of the Road" chapter and the "Signals" chapter in the Bluejackets' manual.

In the first place, however, this difference approaches significance only for the control group (which was not shown the films). In the second place, with the exception of Company 202, only a negligible proportion (not more than four per cent) of any experimental company claimed previous knowledge. In the third place, in Company 202, whose mean test score was higher than that for any other company, those who claimed previous knowledge as here defined actually had an (insignificantly) lower mean score than those who did not

TABLE 15

MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR RULES OF THE NAUTICAL ROAD SERIES:

TOTAL SCORE AND SUBTEST SCORES

Method	Company	n	Total Test			Part 1			Part 2			Part 3		
			Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm
Control	209	84	12.87	3.05	.33	4.39	1.98	.22	4.26	1.59	.13	4.21	1.82	.20
1-Part	202	109	17.38	5.12	.49	6.08	2.54	.24	4.70	2.07	.20	6.60	2.47	.24
	205	84	15.42	4.13	.45	5.55	2.12	.23	4.48	3.04	.19	5.39	2.15	.24
Total		193	16.52	4.81	.35	5.85	2.39	.17	4.60	3.75	.14	6.07	2.41	.17
3-Part	210	80	15.90	3.75	.42	5.85	2.03	.23	4.20	1.91	.21	5.85	2.32	.26
	212	79	15.57	4.67	.53	5.58	2.49	.28	4.30	1.89	.21	5.68	2.35	.27
Total		159	15.74	4.24	.34	5.72	2.27	.18	4.25	1.90	.15	5.77	2.34	.19
Total Experimental Group		352	16.17	4.58	.24	5.79	2.34	.13	4.44	3.73	.10	5.93	2.38	.13

TABLE 16

THE EFFECT OF PREVIOUS KNOWLEDGE ON TEST PERFORMANCE, RULES OF THE NAUTICAL ROAD SERIES: TOTAL TEST SCORE MEANS, STANDARD DEVIATIONS, AND MEAN DIFFERENCES

Method	Company	No Previous Knowledge			Previous Knowledge			Mean Difference	t-Ratio	P
		n	Mean	SD	n	Mean	SD			
Control	209	80	12.70	3.00	4	16.25	1.77	+ 3.55	2.32	< 5%
1-Part	202	88	17.39	5.13	21	17.33	5.07	- .06	.04	> 90%
	205	82	15.35	4.01	2	18.00	7.00	+ 2.65	.89	> 30%
Total		170	16.40	4.73	23	17.39	5.26	+ .99	.92	> 30%
3-Part	210	77	15.75	3.75	3	19.66	.46	+ 3.91	1.77	> 5%
	212	78	15.55	4.70	1	17.00	0	+ 1.45	.30	> 70%
Total		155	15.65	4.26	4	19.00	1.22	+ 3.35	1.56	> 10%
Total Experimental Group		325	16.05	4.53	27	17.63	4.92	+ 1.58	1.73	> 5%

claim previous knowledge. Furthermore, in this company alone, an appreciable proportion (24 per cent) indicated that they had read the chapter in question. It is pertinent to observe, in this connection, that no member of this company reported having seen any of these films before.

One possible explanation for this situation may be that the three-quarters of the company who reported not having read the Bluejackets' manual may have misunderstood the question or deliberately answered it untruthfully. Another might be that those who studied the chapters communicated the results of their study to the rest of the company. A third might be that those who claimed to have read the chapter had not done so. Since relevant data were not available until after the experiment was completed, it is not possible to answer the question.

In any event, since the "previous knowledge" variable did not result in significant intra-company differences, and affected only a negligible proportion of the men, it was not incorporated in further analyses of the data for this film series.

Comparison of the presentation methods. Interpretation of the results obtained when the methods are compared must be made with caution. While no consistent statistically significant effect has been noted for differences in previous knowledge on the part of the subjects, this variable cannot but have increased the errors of measurement and probably, through some differential effect on the groups, made the groups less than strictly comparable with respect to the experimental variable.

The results are, however, consistent with those from the other film series. Table 17 presents the F-ratios for the analysis of variance for the total test score, and for the subtest scores. The F-ratios indicate that the differences between the experimental methods may be attributed to chance fluctuation. In each case, except for the subtest on the second reel, the difference between the experimental and control groups is significant beyond the 0.1 per cent level of confidence. The significant variability among the companies, with respect to the total test score, may be attributed entirely to the performance of Company 202. The mean score of this company is significantly higher than that for any other company.[1]

---

[1] Comparison of the mean of this company with the means of the other three experimental companies yielded, in each case, a t-ratio significant beyond the one per cent level of confidence.

TABLE 17

F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS: RULES OF THE NAUTICAL ROAD SERIES TEST

	Total Test Score	Part 1 Score	Part 2 Score	Part 3 Score
Between experimental and control groups	39.416***	25.461***	< 1	38.414***
Between companies within experimental methods	4.577*	1.503	< 1	6.300**
Between experimental methods	2.498	<1	3.000	1.391

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

*** Significant at the 0.1% level of confidence.

The variance estimates were adjusted for differences in initial status as measured by the Navy General Classification Test and the Mechanical Aptitude Test. The multiple correlation for the experimental group as a whole is .42.

When the variance estimates are adjusted for this correlation, there is a slight reduction in the estimate of error, but in the case of the comparison of the companies there is a proportionately greater reduction in the estimate of the "between companies" variance. The unadjusted "between companies" variance was significant at the one per cent level of confidence; when account is taken of initial status as here defined, the level of significance of the differences among the companies on the film test drops to the five per cent level. It is suggested that the high mean score for Company 202 was at least in part a function of a higher initial status on the matching variables considered jointly.[2]

The interest ratings. It was suggested, when the films were described, that the Rules of the Nautical Road Series involved rather uninteresting material that could be learned only by rote. The arbitrary conventions governing the display of lights and the use of whistle and other signals are not easily integrated into a meaningful whole.

A few of the subjects were interviewed with respect to their interest in the films, and the general impression was received that they did find the films dull and uninteresting. The distributions of rating scores were, for every session, definitely skewed toward the low end of the scale, the modal score being four. However, the scores were distributed over almost the whole continuum.[3]

Means and standard deviations were calculated for the ratings, as well as correlations between the ratings and the test scores (Table 18). For the three-part method companies and group, these are correlations between the rating and the score on the subtest covering the reel shown during the session when the rating was made. For the one-part method companies and group the correlations are between the total test scores and the ratings.

The following observations seem pertinent. First, the means differ inconsequentially. Although many of the differences are highly significant statistically, no reliable meaning can be attached, in a scale as rough as this one, to a difference that is a fraction of a point. **Furthermore,**

---

[2] See footnote 3, Page 27.

[3] Two questions on the adequacy of viewing conditions were included on the rating form, but they were not counted in the rating score. Analysis of these two questions revealed that, although some subjects had poor vision of the screen, the inadequacies in the projection situation were not reflected in test performance.

TABLE 18

MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: RULES OF THE NAUTICAL ROAD SERIES

Method	Film Period	Company	n	Mean Rating	SD	r
1-Part	First (Only)	202	109	3.74	1.35	- .028
		205	84	3.30	1.49	- .288**
		Total	193	3.55	1.43	.157
3-Part	First	210	80	3.87	1.34	.120
		212	79	3.11	1.41	.091
		Total	159	3.50	1.43	.138
	Second	210	80	3.52	1.46	- .091
		212	79	3.72	1.32	.164
		Total	159	3.62	1.39	.033
	Third	210	80	4.39	1.27	.120
		212	79	3.35	1.35	.051
		Total	159	3.87	1.41	.114

** Significant at the 1% level of confidence.

statistically significant intra-method differences (e.g., between companies 210 and 212 for the first and third sessions) are about as frequent as inter-method and inter-sessions differences.

Second, "interest," as measured, is uncorrelated with test performance. All the correlations for this film series (as well as for the other film series) are about zero.

The final analysis with respect to the ratings concerns the distribution of responses to the specific question asked.

Unfortunately, the one question that provided interesting data in the psychology replication - "Do you think this film is too long, too short, or the right length?" - was not included in this briefer questionnaire. The remaining questions tell us little that would serve to differentiate the methods groups. In most cases intra-method and intra-session differences are as great as, or greater than, inter-method or inter-session differences. About 60 per cent at each session (but one) said they would "like to see more films in the series." Over 60 per cent at all sessions (but one) claimed an interest in the subject covered, and over 70 per cent at all sessions (but one) said they "liked the film more or less."

However, these ratings cannot be taken as a very reliable index of interest or attitude. A striking example of the inconsistency that characterizes these ratings is furnished by the responses of Company 212 for the last reel shown them. Approximately 84 per cent of the company reported difficulty in seeing or hearing, and slightly over 50 per cent reported difficulty in both seeing and hearing. Notwithstanding, 81 per cent of the group reported that they thought they "learned something." On the other hand, a majority voted against being shown any more films in the series.

Observation of and talking with samples of the subjects suggested that it would have been difficult, if not impossible, to obtain a rating evaluation of the methods from them, even if the specific question concerning length had been included. In the first place, no adequate frame of reference was available within which they could make the required judgements. Secondly, and more importantly, the experimental situation, whatever its content, afforded a period for relative relaxation and even sleep. This permissiveness was undoubtedly reflected, in part at least, in the ratings, and may be considered to have helped skew the rating distributions toward the low end of the scale. Finally, it was evident at every session that the films, taken singly or together, aroused very little interest in the trainees. The ratings probably reflect to a great extent miscellaneous non-relevant factors in the experimental situation, such as the heat of the day, social pressures implied or actual, and so forth.

Elementary Hydraulics Series

The analysis for the Elementary Hydraulics Series follows closely in approach and conclusions that for the Rules of the Nautical Road Series.

The basic data for the Elementary Hydraulics Series, for the companies and methods groups, are presented in Table 19. Inspection of the table indicates that, with the exception of Company 203, the experimental companies' performances were remarkably homogeneous, differing by only a fraction of a score point either with respect to the total score or to the scores on the three subtests. The mean difference between the experimental group and the control group was five points.

Preliminary to the analysis of the methods differences per se, the effect of previous knowledge of the subject matter was studied. The results with respect to this variable will be reported first.

Previous knowledge. For this film series, "previous knowledge" was defined as a high school course in physics, or a course in hydraulics.

Analysis of the resulting distribution of the subjects' performance, based on this unvalidated and possibly not very reliable criterion, yielded highly significant results. These results are reported in Table 20. The following observations are pertinent: first, between a third and a half of the subjects in each company and methods group reported having taken at least one such course; second, within every company there was a highly significant difference of from 4.2 to 9.8 score points between those who had, and those who had not, had such a course; third, the net effect of "previous knowledge" was such that those members of the control company who reported "previous knowledge" earned a mean score (17.71) that was higher than the mean score of the "no previous knowledge" subgroup in the experimental group (17.65).

This preliminary examination of the data indicated that the comparison between the methods had to take into account the "previous knowledge" variable. Accordingly, it was entered into the variance analysis.

The findings **reported above** suggest, in addition to the evident impact on performance of "previous knowledge," some interaction between the methods and "previous knowledge." In point of fact, such interaction is demonstrated in the following analysis when the experimental and control groups are compared. However, it does not reach a level of statistical significance when the presentation methods are compared among themselves.

TABLE 19

MEANS, STANDARD DEVIATIONS, AND STANDARD ERRORS OF MEANS FOR ELEMENTARY HYDRAULICS SERIES:

TOTAL TEST SCORE AND SUBTEST SCORES

Method	Company	n	Total Test			Part 1			Part 2			Part 3		
			Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm	Mean	SD	SEm
Control	208	109	14.72	4.81	.46	5.07	2.43	.23	5.02	2.05	.14	4.62	1.97	.19
1-Part	203	114	18.92	7.02	.66	5.39	2.89	.27	7.05	2.74	.15	6.47	2.75	.26
	204	76	20.25	7.56	.87	6.21	3.13	.36	7.38	2.67	.19	6.66	3.09	.36
Total		190	19.45	7.27	.53	5.72	3.02	.22	7.18	2.72	.12	6.55	2.89	.21
3-Part	206	59	20.02	5.95	.78	5.58	2.39	.31	7.63	2.38	.20	6.81	2.94	.39
	211	93	20.14	6.29	.66	5.68	2.76	.17	7.58	2.55	.27	6.88	2.57	.27
Total		152	20.09	6.16	.50	5.54	2.77	.25	7.60	2.49	.13	6.85	2.72	.22
Total Experimental Group		342	19.74	6.81	.37	5.68	2.85	.16	7.37	2.27	.09	6.68	2.82	.15

TABLE 20

THE EFFECT OF PREVIOUS KNOWLEDGE ON TEST PERFORMANCE, ELEMENTARY HYDRAULICS SERIES: TOTAL TEST SCORE MEANS, STANDARD DEVIATIONS, AND MEAN DIFFERENCES

		No Previous Knowledge			Previous Knowledge			Mean	t-	P
Method	Company	n	Mean	SD	n	Mean	SD	Difference	Ratio	
Control	208	78	13.53	3.73	31	17.71	5.79	4.18	5.21	< 0.01%
1-Part	203	87	16.61	4.82	27	26.37	7.76	9.76	9.17	< 0.01%
	204	51	17.75	5.99	25	25.36	7.85	7.61	4.62	< 0.01%
Total		138	17.03	5.31	52	28.95	7.82	8.86	8.87	< 0.01%
3-Part	206	39	18.13	5.38	20	23.70	5.25	5.57	3.73	< 0.01%
	211	72	18.58	5.68	21	25.48	5.31	6.90	4.91	< 0.01%
Total		111	18.42	5.58	41	24.03	6.47	5.61	4.86	< 0.01%
Total Experimental Group		249	17.65	5.48	93	25.05	7.31	7.40	10.14	< 0.01%

Comparison of the presentation methods. The results from this film series are consistent with those from the other three film series, and the findings with respect to the effect of "previous knowledge" do not in general affect the basic conclusions.

Table 21 presents the F-ratios for analysis of variance for the total test score and the subtest scores. The "previous knowledge" variable was included only in the analysis of the total score.

In every case, for the total test and for the three subtests, the differences among the companies and those among the methods are insignificant (no F-ratio reaches the five per cent level of confidence). In every case the difference between the experimental group and the control group is highly significant, the F-ratio exceeding the 0.1 per cent level of significance in all instances except for the subtest on the first reel, where it was significant at the five per cent level of confidence.

Table 21 confirms the conclusion that the differences with respect to "previous knowledge" were highly significant, as measured by the total test score. However, among the experimental groups there is no reliable evidence of interaction. That is, although the raw scores indicate a slight reversal in effect (the massed presentation, "no previous knowledge," subgroup has a lower mean than the spaced "no previous knowledge" presentation subgroup, while the reverse is true for the "previous knowledge" subgroups), this reversal may be attributed to chance fluctuation.

When the experimental and control groups are compared, the interaction is significant only at the five per cent level of confidence.

The variance estimates were adjusted for initial status as measured by the Navy General Classification Test and the Mechanical Aptitude Test. When the covariance adjustment[4] is thus made for the total test score, the F-ratio for the "between companies" test and the "between methods" test both increase, in the former case to reach the one per cent level of significance.

This finding suggests that the companies differed somewhat in initial status with respect to the two matching variables considered jointly (they did not differ significantly with respect to these variables considered independently).

---

[4] See footnote 3, Page 27.

TABLE 21

F-RATIOS INDICATING DEGREE OF HETEROGENEITY AMONG THE TEST SCORE MEANS:

ELEMENTARY HYDRAULICS SERIES TEST

	Total Test Score	Part 1 Score	Part 2 Score	Part 3 Score
Between experimental and control groups	63.836***	3.966*	72.783***	50.123***
Between companies within experimental methods	< 1	1.902	<1	<1
Between experimental methods	1.049	<1	2.474	<1
Between previous knowledge and no previous knowledge, experimental companies only	117.977***			
Interaction: experimental methods X previous knowledge	3.045			
Between previous knowledge and no previous knowledge, all companies	108.661***			
Interaction: experimental and control groups X previous knowledge	5.661*			

* Significant at the 5% level of confidence.

*** Significant at the 0.1% level of confidence.

When these differences are taken into account, significant inter-company differences on the film test are indicated. However, the methods groups do not differ significantly even after adjustment. The significant F-ratio should be interpreted as indicating significant intra-method differences rather than inter-method differences. This interpretation is sustained by examination of the means for the film test total scores in Table 19. Three of the experimental company means are about 20, while the fourth, that for Company 203, is 18.9.

The interest ratings. The significance and the limitations of the interest ratings collected during the Navy replication were discussed in some detail earlier in this chapter in connection with the findings for the Rules of the Nautical Road Series. Although the Elementary Hydraulics Series seemed to win a slightly greater degree of interest than the preceding series, the comments previously made apply equally here. Therefore, they will not be repeated.

The findings specific to this series are presented in Table 22, which embodies **the same kind of information as was given for the Rules Series.**

The ratings for the Elementary Hydraulics Series are also skewed toward the low end of the scale. Again, the modal rating was four, except for the first session for Company 211, where it was two (close to the "unfavorable" end of the scale).

The ratings are uncorrelated with test performance, and intra-session mean differences are generally greater than inter-session or inter-method differences. The mean rating for the second and third sessions of the spaced method are significantly higher than the mean rating for the massed method. However, the mean rating for Company 204 (massed presentation) is higher than for all but the means for the second and third sessions for Company 206. Basically, whatever the statistical level of significance, these differences cannot be considered as other than inconsequential.

The final analysis of the ratings pertains to the distribution of responses to the specific questions asked. Here again, intra-session variability is generally greater than inter-session or inter-method variability. However, in several of the questions, at least indirect evidence is afforded of a slightly better reception for the spaced method. In general, a larger proportion of the subjects attending each session of the spaced method presentations, as compared with the subjects attending the massed presentations, indicated that they would "like to see more films in the series," and that they learned "something" or "a great deal" from the films. In response to every question but one, Company 203, which participated in the massed method presentation, afforded the largest proportion of negative responses (59.6 per cent would not like to see more

TABLE 22

MEAN INTEREST RATINGS, AND CORRELATIONS BETWEEN RATINGS AND TEST SCORES: ELEMENTARY HYDRAULICS SERIES

Method	Film Period	Company	n	Mean Rating	SD	r
1-Part	First (Only)	203	114	2.86	1.54	.248**
		204	76	3.85	1.60	- .235*
		Total	190	3.26	1.64	.086
3-Part	First	206	59	3.41	1.48	.098
		211	93	3.47	1.53	.156
		Total	152	3.45	1.51	.170
	Second	206	59	4.15	1.41	.089
		211	93	3.81	1.38	.171
		Total	152	3.94	1.40	.164*
	Third	206	59	4.39	1.27	.120
		211	93	3.35	1.35	.051
		Total	152	3.87	1.41	.114

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

films in the series, 54.4 per cent were not interested in the subject matter). In the one other question, "Did you learn from this film?", 7.9 per cent of the company said "nothing;" in its third session, 10.2 per cent of Company 206 made the same response.

Comparisons of these results with the results from the Rules of the Nautical Road population are practically meaningless, especially so in view of the fact that no consistent trend is evident. One small point may be noted, however. The incidence of "I learned nothing" responses was, for the sample as a whole, generally greater for the Hydraulics Series than for the Rules Series. This observation, which is not validated by a meaningful significance statistic, is consistent at least with the general character of the situation.

The Elementary Hydraulics Series developed concepts for which most of the subjects could claim at least the familiarity of recognition. The Rules Series, on the other hand, presented material that was almost completely unfamiliar, and if it was accepted at all the subject felt he acquired something new.

VI. COMBINED RESULTS AND DISCUSSION

The specific statistical data for the two experiments have been presented in considerable detail in the previous chapters. The data will therefore not be repeated, except where essential, in this summary of the results from the four film series.

It is convenient to summarize the results in two categories: those applicable to the four series, and those applicable to only one or two series. The former will be treated first.

General Results

The following are the significant general results. They obtain, except as specifically qualified, for all four film series.

1. Comparability of the groups: The control and experimental groups for each film series were substantially equivalent as measured with respect to initial status. The variations among the psychology classes with respect to the two indices, grade-point-average-to-date and final psychology grade, and the variations among each set of Navy companies with respect to the Navy General Classification Test and the Mechanical Aptitude Test may, in both instances, be attributed essentially to chance fluctuation.

2. Reliability of the film tests. The four film tests were of varying reliability as measured by the Kuder-Richardson technique of rational equivalence. For the total experimental group, these reliabilities were: Elementary Hydraulics test, .80; Rules of the Nautical Road test, .57; Ape and Child test, .51; and Cat Neurosis test, .73.

3. Measure of absolute learning. "Absolute learning" may be defined as the difference between the test performance of a control group not exposed to the experimental stimulus (the film series) and that of an experimental group exposed to this stimulus. For all four series, a highly significant difference was obtained as between the experimental and control groups. However, in terms of the tests the Navy groups seem to have learned relatively much less than the psychology classes. These results may be a function of lack of interest and motivation in the Navy population, of the difficulty of the material or the tests, of the inadequacy of the projection facilities, or of other unidentified factors.

It should be noted that, while objective evidence on the point is lacking, it was the consensus of those who viewed them that the *Cat Neurosis* films are much more difficult than the *Ape and Child Series*. The *Rules of the Nautical Road* films, similarly, made considerably greater demands on the learner than the *Elementary Hydraulics Series*.

4. *Inter-methods differences*. The psychology classes, it will be remembered, were shown the series in three spacing patterns: a massed one-hour session, two spaced 30-minute sessions, and four spaced 15-minute sessions. The Navy companies were shown the series in two patterns: a massed 45-minute session and three spaced 15-minute sessions. Except for three classes for the second psychology series, all groups were tested one week (approximately) after the experimental sessions.

For all four series, the *inter-methods* differences in *mean total test score*, at the end of one *or* two weeks, may be attributed to chance fluctuations.

Furthermore, for all series but one (*Rules of the Nautical Road*) the *inter-company* or *inter-class* differences in total score may be attributed to chance fluctuation. For the *Rules of the Nautical Road Series*, a significant inter-company difference was noted. This resulted from a difference between the performances of the two companies exposed to the massed presentation. In general, there was *a tendency* for the means of the spaced methods groups to be higher than the means for the massed method groups. This tendency was not consistently present, however; the discrepant company mean in the massed method group for the *Rules of the Nautical Road Series* was higher than the means for the companies in the spaced group.

In general, these results apply equally to the subtest scores. However, for one subtest for the *Cat Neurosis Series* and for two subtests for the *Ape and Child Series*, the analysis of variance indicated variation greater than could be accounted for by chance factors alone. In one of the *Ape and Child* subtests the mean difference was in favor of the massed presentation; in the other two instances it was in favor of the spaced presentations. Since these results are based on tests including only 20 items, and since the bulk of the evidence is in the other direction (namely, that no statistically significant methods differences exist), they cannot be considered as seriously disturbing the stability of the principal conclusion.

5. *The effect of initial status*. It has already been demonstrated that the groups were essentially homogeneous with respect to the "initial status" or matching variables, considered individually. The question remained, when these variables (grade-point average and final psychology grade, or

General Classification Test and Mechanical Aptitude Test) are used jointly to define initial status, and the groups are matched by means of an analysis of covariance, do significant methods differences then emerge? It was noted, first, that except for the Elementary Hydraulics test, the combined matching variable correlated to only a moderate or low degree with the film test score. Second, in no case did adjustment for "initial status," as defined, result in indicating significant methods differences.

6. The interest ratings. Three observations are pertinent to the interest rating results. First, in general, intra-session variation and inter-company variation were as great as inter-method variation, and the rating scores did not seem to differentiate among the methods. For the Cat Neurosis Series, however, the mean rating scores of the massed method group were consistently and significantly lower than the means for the spaced groups. Second, analysis of the distribution of responses to most of the specific questions failed to provide any consistent indication of methods differences. It was suggested, however, that the question "Was this film too short, the right length, or too long?", which was asked of the psychology classes only, did differentiate. Subjects shown the series in the massed presentation method were the only ones who reported the session "too long." In the case of the Ape and Child Series, only small percentages of the subjects made this response; in the case of the Cat Neurosis Series, this response was made by 60 to 80 per cent of the subjects. Third, the correlation between interest rating score and test score was about zero.

The data suggest that "interest" ratings reflect factors other than the adequacy of the films as teaching devices.

Specific Results

Two subsidiary questions were investigated in this study. These relate to (1) the effect of lengthening the retention period from one to two weeks, and (2) the effect of previous knowledge in relation to the presentation methods.

1. One week versus two week retention periods. Although it had been planned to test each class one week after the experimental sessions, certain considerations beyond the experimenter's control made it necessary to test one class in each methods group in the psychology classes' experiment, for the Cat Neurosis Series only, two weeks later. Conclusions based on this single instance cannot be considered too stable,

but they are suggestive. First, it was noted that extension of the retention period was accompanied by relatively lower scores for all the experimental classes involved (in comparison with the one-week group). Second, the spread of means for the methods groups was somewhat greater for the two-week test than for the one-week test, with the massed group earning the lowest score. Finally, however, the inter-methods differences were statistically insignificant.

2. Previous knowledge. In the Navy replication it seemed likely that the subjects would be rather heterogeneous with respect to previous educational background, and that this might affect the results for the Elementary Hydraulics Series. Men who had had high school physics would have a sensible advantage over those who had not. It seemed possible, furthermore, that, although the relevant sections in the Navy Apprentice Seaman handbook, the Bluejackets' manual, dealing with visual and whistle signals are usually not covered at an early a period in training, some of the men might have read the relevant sections anyway. There was also the possibility that an enterprising company commander had assigned them ahead of time.

Therefore, the subjects seeing each series were asked to indicate whether they had "previous knowledge" of the subject, as defined above.

For the Rules of the Nautical Road Series, it was found that, except in one company, only a neglible number (one to four) of the men claimed such knowledge. However, the differences between the "previous knowledge" and "no previous knowledge" subgroups in each company were negligible and statistically insignificant. This was particularly true for the company in which the largest proportion of the men claimed "previous knowledge."

For the Elementary Hydraulics Series, on the other hand, it was found that "previous knowledge" was a very potent variable. In every company a large and highly significant difference was found between those who had, and those who had not, had a course in high school physics or its equivalent. The mean of the "previous knowledge" subgroups in the control group was, in fact, higher than the mean of the "no previous knowledge" experimental group.

Nevertheless, analysis revealed that there were no significant inter-methods differences within either knowledge category, and that "previous knowledge" did not affect the relative effectiveness of the methods.

VII. CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The results of the investigation seem to support the following general consluslons.

1. When a typical hour-long series of instructional motion pictures is used as the sole teaching tool, students learn about the same amount from the series whether they are shown all the reels comprising the series in one long training session, or one reel at a time in several short training sessions.

2. Increasing the length of the training session to one hour does not seem to result in a diminution of interest on the part of the learners. Furthermore, a learner's test performance is practically independent of his rated interest in the films.

3. Long massed film sessions are about as effective in ensuring long-term (two-week) retention of the film content as short spaced sessions are.

4. While previous knowledge of some parts of the film content results in higher test scores, the effects of previous knowledge or its lack are about the same whether the reels in a training films series are presented in one long, or in several short spaced, sessions.

The general consluclon may be stated as follows: that a few hour-long film training sessions, like hour-long classes, result in learning about as efficient as that achieved by many short training sessions.

Applications and Recommendations

The extent to which the findings presented are applicable to particular training programs depends primarily upon two considerations. These are, first, the question of whether the training authorities are inclined to use films as aids or as total teaching tools, and second, the question of how long training sessions themselves are to last. If training sessions are to be only an hour long, and films are to be relegated to the position of aids considerably supplemented by other class work, then there is simply not the time in any one period for both a "long" film and other work. Of course, one alternative might be to use the preceding and following periods for the preparatory and follow-up work. In that case these findings suggest, not that an hour is the limiting length of an effective teaching film, but that an hour film is about as satisfactory as several shorter spaced films (which add up to the hour film).

On the other hand, if primary dependence is to be placed on the film as a "total teacher" then these findings afford justification for relatively long film training sessions. Research on instructional films, to which reference has been made, suggests that it is entirely feasible, for at least certain contents, to use films as the exclusive means of instruction. This study has, in general, suggested that films-as-teachers can be considered as having the same scheduling requirements as human lecturers.

Following are specific recommendations for application of these results.

1. In mass training programs the scheduling of long film sessions for training purposes should be explored as a means of economizing training time, simplifying scheduling, and utilizing instructors more efficiently.

2. Producers should consider the possible advantages of making single long films, where the material calls for extended treatment, rather than making series of short units. It is believed that in this way significant economies in production may be realized, as well as better integration and more consistent treatment of the material.

3. Further research is needed to determine what are the limits in lengthening film sessions, what kinds of subject matter can be taught most efficiently in concentrated sessions, what film production techniques are most appropriate to long training films, and whether long film sessions are as effective as short sessions with all or most learner populations.

There are certain other questions that merit investigation. First, to what extent is subject-matter difficulty related to the time variable? Second, what would be the effect of varying the intervals between the sessions? Perhaps spacing the films one or two hours apart on the same day, rather than one or two days apart, would yield appreciable advantages over a massed presention. Third, what are the conditions (nature of the film, its quality, adequacy of facilities, and so forth) that lead to the judgment that one film lasting an hour is too long, and another of the same length is not?

Finally, if these findings are to be meaningful, learning theory must be extended to include the appropriate situations. It has been suggested that the findings with respect to massed versus distributed practice and whole versus part learning have at best a spurious face validity in relation to educational practices in general. These concepts fail to provide a rubric under which either the conduct of ordinary class work or film instruction such as is used here can be easily subsumed. Some of the conclusions reached in the educational research literature cited are, if anything, contrary to principles adduced from experiments in these two traditional learning categories. It is suggested that both experimentation and theory applicable to relatively complex, meaningful, highly structured materials are needed.

ACKNOWLEDGEMENTS

A study of the sort reported here in never the work of a single individual. The writer is under numerous and heavy obligations for advice and assistance, for which grateful acknowledgement is made.

Dr. Clarence R. Carpenter, Director of the Instructional Film Research Program, suggested the initial idea out of which this study was formulated, and gave valuable advice and criticism during the course of its development. Dr. Kinsley R. Smith, Professor of Psychology, served as Chairman of the author's thesis advisory committee, and guided the research; Dr. Albert K. Kurtz, Dr. Bruce V. Moore, Dr. Joseph DeCamp and Dr. Carpenter, all members of the thesis advisory committee, reviewed the final manuscript and made many helpful suggestions. Mr. George Bender and Miss Jean MaGuire assisted in constructing the tests used. Mr. John Kishler aided in conducting the Navy replication, and Mrs. Gloria Kahn assisted in the statistical calculations.

Captain F. J. Grandfield, USN, Commanding Officer, Recruit Training Command, Great Lakes Naval Training Center, Great Lakes, Illinois, granted permission for testing at the Center. Lt. (j.g.) J. M. Bauer, USN, Scheduling Officer of the RTC, selected and scheduled the companies of recruits used.

Dr. Kendon R. Smith, Research Coordinator of the Instructional Film Research Program rendered valuable assistance in the editing of this report.

REFERENCES

1. Bernard, E. G. Silent films and lantern slides in teaching French. *Modern Language*, 1936, *21*, 109-115 (November).

2. Borow, H. *A statistical analysis of the predictive measures of freshman academic achievement in use at The Pennsylvania State College*. Unpublished master's thesis, The Pennsylvania State College, State College, Pa., 1942.

3. Bruns, E. L. Experimental high-school program. *Dept. Sec. Schl. Princ. B.*, 1939, *23*, 41-42 (May).

4. Calo, G. Cinema and teaching methods. *Int. Rev. educ. Cinematography*, 1934, *6*, 353-358.

5. Castore, G. F. *A screening and selection battery for prospective physicists and chemical engineers*. Unpublished doctoral dissertation, The Pennsylvania State College, State College, Pa., 1948.

6. Clevenger, A. W. Long period daily class schedule for high schools. *North Central Ass. Quart.*, 1936, *10*, 456-461.

7. Coblentz, I. *Prognosis of freshman academic achievement at The Pennsylvania State College*. Unpublished doctoral dissertation, The Pennsylvania State College, State College, Pa., 1943.

8. Dale, E., Dunn, F. W., Hoban, C. F., Jr., and Schneider, E. *Motion pictures in education*. New York: H. W. Wilson Company, 1938.

9. Devereux, F. L. *The educational talking picture*. Chicago: The University of Chicago Press, 1933.

10. Doane, D. C. What makes a good educational film? *Educ. Screen*, 1939, *15*, 203-206, 239-241, 271-273, 305-307.

11. Greenley, K. F. Single periods vs. double periods. *Sch. Exec.*, 1943, *63*, 34-35 (December).

12. Hall, W. E., and Cushing, J. R. The relative value of three methods of presenting learning material. *J. Psychol.*, 1947, *24*, 157-162.

13. Hoban, C. F., Jr. *Focus on learning*. Washington: American Council on Education, 1942.

14. Jayne, C. D. A study of the learning and retention of materials presented by lecture and by silent film. J. educ. Res., 1944, 38, 47-58.

15. Kambly, P. E. Comparison of a 1-hour two-semester and a 2-hour one-semester course in biology. Sch. Sci. and Math., 1939, 39, 279-281.

16. Kuder, G. F., and Richardson, M. W. The theory of estimation of test reliability. Psychometrika, 1937, 2, 151-160.

17. McGeoch, J. A. Whole-part problem. Psychol. Bull., 1931, 28, 713-739.

18. McGeoch, J. A. A reevaluation of the whole-part problem in learning. J. educ. Res., 1932, 26, 1-5.

19. McGeoch, J. A. The whole-part problem in memorizing poetry. J. genet. Psychol., 1933, 43, 439-447.

20. McGeoch, J. A. The psychology of human learning. New York: Longmans, Green and Co., 1942.

21. McKnown, H. C., and Roberts, A. B. Audio-visual aid to instruction. New York: McGraw-Hill, 1940.

22. McMillin, J. B. Economy and the hour period. Amer. Sch. Bd. J., 1934, 88, 50.

23. Manheimer, W. A. Suggested experiment of the reorganization of the high school on the basis of a longer recitation period. High Points, 1937, 19, 18-26 (January).

24. Mertens, M. S. Prediction of academic achievement for sophomore women in the liberal arts and education curriculums. Unpublished master's thesis, The Pennsylvania State College, State College, Pa. 1948

25. Nord, G. E. Our school tested six and eight period days. Clearing House, 1941 , 16, 108-110 (October).

26. Pennsylvania State College. Psychological Cinema Register (No. Six, 1949 and Forward). State College, Pa.: The Pennsylvania State College, 1949.

27. Philpott, S. J. F. Cinema Commission of Inquiry experiment; a discussion of results and experimental methods. Brit. J. educ. Psychol., 1946, 16, 32-38.

28. Roulette, K. K. The validity of The Pennsylvania State College Academic Aptitude Examination for predicting the academic success of sophomore men in the school of

liberal arts at The Pennsylvania State College. Unpublished master's thesis, The Pennsylvania State College, State College, Pa., 1948.

29. Schultz, D. G. *The efficiency of The Pennsylvania State College Psychological Examination in the prediction of freshmen academic achievement*. Unpublished master's thesis, The Pennsylvania State College, State College, Pa., 1942.

30. Stewart, H. H. *A comparative study of the concentration and regular plans of organization in the senior high school*. Teachers College Contributions to Education No. 600. New York: Columbia University Press, 1934.

31. United States Navy. *The Bluejacket's manual* (Thirteenth edition). Annapolis, Md.: United States Naval Institute, 1946.

32. United States Navy Department. *Catalog of training films for the United States Navy and Marine Corps*. Compiled by the Motion Picture Productions Section, U. S. Naval Photographic Service. Washington, D. C.: U. S. Navy Department, July 1946. (Restricted).

33. U. S. War Department. *How long should training sessions be*? Morale Services Division, Army Service Forces. (Photostat, unpublished).

34. VanderMeer, A. W. Relative effectiveness of exclusive film instruction, films plus study guides and typical instructional methods. Preliminary report of results for Project No. 15. In: *Instructional Film Research Program Progress Report No. 10*, Period 1 January to 28 February 1949. State College, Pa: The Pennsylvania State College, Instructional Film Research Program, 1949.

35. Weaver, P. C. *Scholastic ability and progress in college in relation to five high school factors*. Doctoral dissertation, Columbia University, New York, 1935.

36. Whittaker, B. A. *A validation study of the prognosis of academic success at The Pennsylvania State College*. Unpublished master's thesis, The Pennsylvania State College, State College, Pa., 1943.

37. Woodworth, R. S. *Experimental psychology*. New York: Henry Holt, 1938.